5 Neurophilosophy

Georg Northoff

5.1 Background: The History of Neurophilosophy

Neurophilosophy stands for the investigation of philosophical questions in the context of a neuroscientific hypothesis.

Recent neuroscientific progress has led to the extension of neuroscience to apply and include also concepts like consciousness, free will, self, etc. that were originally discussed in philosophy. This has led to the recent emergence of a new field – neurophilosophy. The term "neurophilosophy" is often used either implicitly or explicitly for the characterization of an investigation of philosophical theories in relation to neuroscientific hypothesis. According to Breidbach [1], pp. 393–394, "neurophilosophy" had already been implicitly practiced at the turn of the last century by W. Wundt (1832–1920), for instance. Another neurophilosopher, though not named as such, was Schopenhauer who was probably the first philosopher to introduce the concept of the brain in the philosophical context. The French philosopher M. Merleau-Ponty (1908–1961) may also be considered a neurophilosopher since in his 'Phenomenology of perception' he explicitly introduces the brain and its neural organisation and links it to perception and other originally philosophical concepts.

Other important developments in this regard were put forward by the American philosopher W. von Orman Quine (1908–2000): He raised the question whether what we can know about ourselves and the world as usually dealt with in the philosophical discipline of epistemology can be traced back to nature itself and ultimately to evolution [2]. This was complemented by the collaboration between the philosopher K. Popper (1902–1994) and the neuroscientist J.C. Eccles (1903–1997) who discussed the relation between brain and mind from both perspectives, neuroscientifically and philosophically [3]. Finally, the term 'neurophilosophy' was explicitly coined by the American philosopher P. Churchland [4] in her book 'Neurophilosophy' where she discussed empirical results side by side with theoretical issues.

The current field of neurophilosophy covers mainly three different domains, 'Empirical Neurophilosophy', 'Practical Neurophilosophy', and 'Theoretical Neurophilosophy'. 'Empirical Neurophilosophy' describes the "application of neuroscientific concepts to traditional philosophical questions" [5], p. 1. Here concepts like consciousness, self, and free will (see below for details) that have traditionally been dealt with theoretically in philosophy are now investigated experimentally in neuroscience. Secondly, there is the field of 'Practical Neurophilosophy' that deals with ethical concepts like free will, moral judgment, and informed consent in the neural context of the brain. Thereby, as in empirical neurophilosophy, the philosophical-ethical concepts may also be extended from the originally purely human domain to animals, like

G. Northoff
Canada Research Chair for Mind, Brain Imaging, and Neuroethics, Michael Smith Chair for Neuroscience and Mental Health, University of Ottawa Institute of Mental Health Research,
Ottawa, Canada
e-mail: georg.northoff@theroyal.ca

C.G. Galizia, P.-M. Lledo (eds.), *Neurosciences - From Molecule to Behavior: a University Textbook*,
DOI 10.1007/978-3-642-10769-6_5,

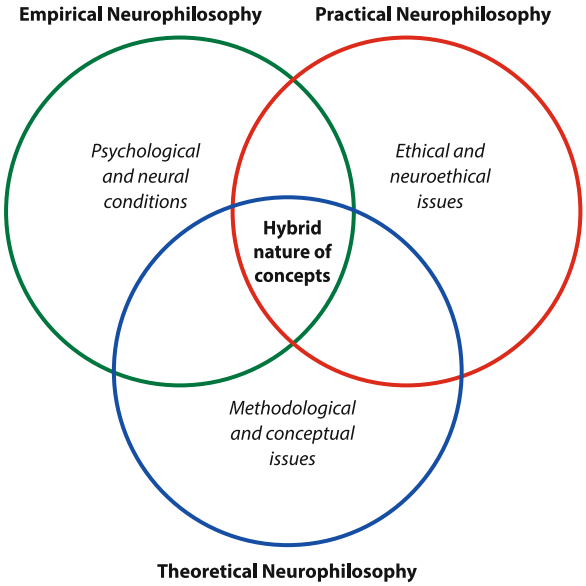

Fig. 5.1 The figure illustrates the three main domains of neurophilosophy, empirical, theoretical, and practical. Empirical neurophilosophy is concerned with the search for the neural and psychological conditions of originally philosophical terms like self, consciousness, free will, etc. Theoretical neurophilosophy is about the methodological and conceptual issues when linking neuroscientific data/facts and philosophical concepts. Finally, practical neurophilosophy is about the linkage between neuroscience and ethics with ethical issues in neuroscience and neuroscientific mechanisms underlying ethical concepts

whether the latter have free will or not. Third, and finally, there is 'Theoretical Neurophilosophy' which focuses on methodological issues like how to link empirical data and theoretical concepts in neurophilosophical investigation (Fig. 5.1).

5.2 Empirical Neurophilosophy – Experimental Investigation of Philosophical Concepts

One of the main originally philosophical concepts investigated in neuroscience is consciousness. What is consciousness? **Consciousness** is often understood as the ability to detect, evaluate and report about the experience of a particular object or event in the environment or the own thoughts. Since detection, reporting, and evaluating requires access to the content in question, this form is often called 'access consciousness'.

'Access consciousness' must be distinguished from the experience itself which, following philosophers like Th. Nagel [6], can be characterized by a particular point of view, a stance in the world, from which we perceive and experience ourselves and others. That form of consciousness has been described as 'phenomenal consciousness'. The distinction between 'phenomenal and access consciousness' is considered by many a core distinction which has also aroused plenty of controvery. Some authors deny for instance that phenomenal consciousness can be distinguished from access consciousness. However, animals may have phenomenal consicousness while they may remain unable to report the contents of their consciousness thus lacking access consciousness.

Following Christoph Koch and Francis Crick [7] we need to identify what they call the 'neural correlates of consciousness' (NCC). The NCC describe the search for those minimally neuronal conditions that are jointly sufficient for any one specific conscious, i.e., phenomenal, percept that we can experience. Several neuronal mechanisms have been discussed as possible candidate mechanisms for the NCC. In the following I highlight some of the main and most popular suggestions.

G. Edelman [8] considers cyclic processing and thus circularity within the brain's neural organisation as central for constituting consciousness. Cyclic processing describes the re-entrance of neural activity in the same region after looping and circulating in so-called re-entrant (or feedback) circuits.

This is for instance the case in primary visual cortex (V1): The initial neural activity in V1 is transferred to higher visual regions such as the inferotemporal cortex (IT) in feedforward connections. From there it is conveyed to the thalamus which relays the information back to V1 and the other cortical regions implying thalamocortical re-entrant connections. Consciousness is assumed to be constituted on the basis of such feedback or re-entrant connections that allow for cyclic processing.

What is the exact mechanism of the feedback or re-entrant circuits? Re-entrant circuits integrate information. This leads Giulio Tononi to emphasize the integration of information as the central neuronal mechanism in yielding consciousness. He consecutively developed what he calls 'Integrated Information Theory' (IIT). We usually focus on the content that is selected to become conscious, i.e., 'what is perceived'. Instead, as the ITT claims, we may better search for the neuronal mechanisms that allow excluding content from becoming conscious, i.e., 'what is ruled out'. The information that is ruled out to become conscious may suffer from insufficient integration of information and remains therefore unconscious.

Tononi assumes the integration of information to be particularly related to the thalamo-cortical re-entrant connections: These re-entrant connections process all kinds of stimuli thus remaining unspecific with regard to the selected content. They make it possible to generate a particular point of view and an associated quality of experience (also called qualia) as hallmark feature of consciousness. Linkage of these qualia to the content processed via thalamo-cortical information integration may then allow these contents to become conscious. This distinguishes them from the unconscious contents that do not undergo such cyclic processing via the thalamus – and therefore the addition of the specific quality, the qualia, remains impossible.

Another suggestion for the neural correlate of consciousness comes from B. Baars [9, 10] and others like S. Dehaene. They assume global distribution of neural activity across many brain regions in a so-called global workspace to be central for yielding consciousness: The information and its contents processed in the brain must be globally distributed across the whole brain in order for them to become associated with consciousness.

When information is only processed locally within a particular region but not throughout the whole brain, it can not be associated with consciousness anymore. The main distinction between unconsciousness and consciousness is thus supposed to be manifest in the difference between local and global distribution of neural activity. Hence, the global distribution of neural activity is here considered a sufficient condition and thus neural correlate of consciousness.

Taken together, there are currently these neuroscientific suggestions for consciousness. Future research is needed though to further specify the neuronal mechanisms themselves and the features of consciousness itself. Consciousness may by itself not be as homogenous as it appears; instead, it may be characterized by different features as for instance a point of view (see above), a quality (see above), and a particular unity as unifying convergence point for different contents.

Another originally philosophical concept now hotly debated in neuroscience is the concept of the *self*. The question of the self has been one of the most salient problems throughout the history of philosophy and more recently also in psychology and neuroscience. For example, William James (1842–1910) distinguished between a physical self, a mental self, and a spiritual self. These distinctions seem to reappear in recent concepts of self as discussed in neuroscience. Damasio [11] and Panksepp [12] suggest a "proto-self" in the sensory and motor domains, respectively, which resembles James' description of the physical self. Similarly, what has been described as "minimal self" [13, 14] or "core or mental self" [11] might correspond more or less to James' concept of mental self. Finally, Damasio's "autobiographical self" and Gallagher's "narrative self" strongly rely on linking past, present, and future events with some resemblances to James' spiritual self.

These distinct selves are now related to distinct brain regions. For instance, the "proto-self" outlining one's body in strongly affective and sensory-motor terms is associated with subcortical regions like the periaqueductal gray, the colliculi, and the tectum. The "core or mental self" building upon the "proto-self" in mental terms is associated more with the thalamus and cortical regions like the ventromedial prefrontal cortex (see, for instance [11, 15]). Finally, the "autobiographical or extended self" that allows one to reflect upon one's "proto-self" and "core or mental self" is associated with cortical regions like the hippocampus and the cingulate cortex.

Humans show various cortical regions, predominantly the so-called cortical midline structures (CMS), to be involved in what is called self-related processing (SRP) that are integrated with subcortical processes to yield an integrated subcortical-cortical midline system (SCMS). The lowest regions of this distributed SCMS network include the periaqueductal gray, the superior colliculi, and the adjacent mesencephalic locomotor region as well as preoptic areas, the hypothalamus, and dorsomedial thalamus, while cortical regions include the ventro- and dorsomedial prefrontal cortex, the pre- and supragenual anterior cingulate cortex and the posterior cingulate cortex, and the medial parietal cortex. The association of the subcortical regions with a sense of self has led to the assumption that already animals may have a sense of self [16, 17] though most likely not as cognitively elaborated as the human self.

5.3 Theoretical Neurophilosophy – Methodology and Knowledge of the Linkage Between Brain Data and Philosophical Concepts

One of the main issues in neurophilosophy is the question for methodology. How can we link empirical data, so-called facts as obtained in neuroscience, to the concepts and their meaning as dealt with in philosophy?

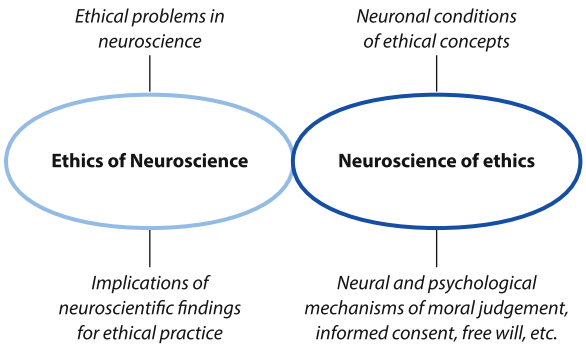

Fig. 5.2 The figure illustrates the fields of practical neurophilosophy. Ethics of neuroscience concersn ethical problems in neuroscience like informed consent. While neuroscience of ethics refers to the neural and psychological conditions of ethical concepts

Data and facts do not require any definition and determination. They rely on observation and can in principle be obtained by anybody; they thus remain investigator-independent. This is different in the domain of concepts. Concepts carry a meaning, a semantic dimension, which may be closely related to the investigator and how he defines and uses the concept in question.

However, empirical-experimental investigation cannot do without concepts. For instance, in formulating the hypotheses of the experiment as starting point for developing an appropriate experimental design, concepts play a substantial role. And after obtaining the data they must be interpreted for which again concepts are necessary. This concerns only concepts within the natural world, the world we live in, and thus what philosophers call the 'natural conditions'. Such 'natural conditions' must be distinguished from 'logical conditions' that describe logically possible worlds which may or may not be realized within the context of our current natural world.

The neurophilosopher is thus confronted with the principal gap between data/facts and concepts in a twofold manner. First, there is the gap between data/facts and concepts within the domain of the natural world: How do certain data about, for instance, the reward system in animals stand to the concept of reward in general in both animals and humans? This is a gap the neuroscientists themselves already face which, due to the predominant experimental focus, is often neglected. Secondly, there is the gap between neuroscientific data/facts in the natural world and the concepts in the logical worlds of the philosophers: How can we infer from neuroscientific data about consciousness to the philosophical concept of consciousness and, vice versa, how can we translate the latter into experimental designs to test it empirically? This is a truly neurophilosophical gap which we need to bridge if neurophilosophy is to succeed in both methodology and knowledge.

5.4 Practical Neurophilosophy – Neuroethics and the Relevance of Ethical Concerns in Neuroscience

Practical neurophilosophy or neuroethics focuses, on the one hand, on the investigation of the psychological and neural conditions of ethical concepts like free will, decision making, moral judgment, and informed consent. This can be described as 'neuroscience of ethics'. At the same time, practical neurophilosophy also deals with ethical problems in neuroscience and thus with issues of validity of informed consent in psychiatric patients, enhancement of cognitive functions by neuroscientific interventions, coincidental findings in neuroimaging (Fig. 5.2). That amounts to an 'ethics of neuroscience' [18].

Do we have a free will or not? The free will is, for instance, manifest in our daily decisions if, for instance, we choose the red rather than the green apples in the supermarket. Recent neuroscience detected the neural mechanisms of decision making that seem to involve a number of different brain regions including those where reward is processed. The reward regions include the ventral tegmental area (VTA), the ventral striatum (VS), and the ventromedial prefrontal cortex (VMPFC). All these regions have their homologs in various animal species so that the same question for the free will may also be extended from human to animals.

The observation that the apparently free decision making is related to and, in fact, temporally preceded by neural activity specifically related to the decision in question has put the concept of free will in doubt. If the free will is pre-determined by the neural events in the brain, one can no longer speak of a free will. The free will is then no longer free but nothing but a mere illusion on our side with the brain determining our actions and decisions. Are we thus no longer free in our will? That obviously is an interpretation of the data and also depends on the definition of the concept of free will. If

one, for instance, presupposes a narrow concept of free will that excludes any preceding changes, the present brain data may tell us that there is indeed no free will. Brain data and free will are then incompatible. Conversely, a wider concept of free will that does not exclude preceding neural activity changes may then be well compatible with the brain data.

The debate about free will pertains to a wider issue, the question of determinism versus indeterminism. Determinism assumes that all our decision and also what we call free will is determined completely and exclusively by the brain and its neural activity. Our person or our self, as presumably distinct from the brain, has then no say at all in our decision. Hence, it is then the brain rather than the self that makes the decision and has a 'neuronal will' rather than a 'free will'. That however is countered by indeterminism. Indeterminism argues that the brain itself and its neural activity changes does not determine completely and exclusively our decision making so that there are traces of free will left in our decisions. Who is right, determinism or indeterminism? As said above, it may strongly depend not only on the data but also on the conceptual definitions.

Besides such questions belonging to the 'neuroscience of ethics', the neuroscientific investigation of ethical concepts, there are also issues pertaining to ethical problems in neuroscience. One problem here is, for instance, the one of informed consent which subjects have to give when participating in experimental investigations. Being able to give informed consent may include a variety of different functions, cognitive, social, and affective, that are all ultimately brain-based. Does this mean that we have to exclude those subjects that suffer from impairments in these functions? Furthermore, recent research demonstrates that animals possess many of the cognitive and social functions originally attributed to humans only. Do we therefore need to develop more rules for animal participation in research by, for instance, considering that they can have consciousness, feel pain, and empathize with co-species?

5.5 Summary

Neurophilosophy is a young and novel field right at the intersection between neuroscience and philosophy. Unlike more established disciplines, it has not yet an established method that needs to be developed in the future as part of a 'theoretical neurophilosophy'. At the same time though neurophilosophy is a highly promising field which will be able to provide novel answers to questions discussed in philosophy for more than 3,000 years. This will not only enrich neuroscience and provide new ideas for experimental designs but will also change and reverberate in philosophy itself by allowing for a shift from the hitherto mind-based philosophy to a more brain-based neurophilosophy.

References

1. Breidbach O (1997) Die Materialisierung des Ichs – Eine Geschichte der Hirnforschung im 19. und 20. Jahrhundert. Suhrkamp, Frankfurt aM
2. Quine WO (1969) Epistemology naturalized. In: Quine WO (ed) Ontological relativity and other essays. Columbia University Press, New York
3. Popper K, Eccles J (1989) Das Ich und sein Gehirn. Piper, München
4. Churchland P (1986) Neurophilosophy: toward a unified science of the mind-brain. MIT Press, Cambridge, MA
5. Bickle J, Mandik P, Landreth A (2006) The philosophy of neuroscience. Journal [serial on the Internet]. Available from: http://plato.stanford.edu/entries/neuroscience/
6. Nagel T (1979) What is it like to be a bat? Mortal questions. Cambridge University Press, New York, p 166
7. Koch C, Crick F (2001) The zombie within. Nature 411:893
8. Edelman GM (2003) Naturalizing consciousness: a theoretical framework. Proc Nat Acad Sci U S A 100:5520–5524
9. Baars BJ (2005) Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Prog Brain Res 150:45–53
10. Baars BJ, Franklin S (2007) An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA. Neural Netw 20:955–961
11. Damasio A (1999) How the brain creates the mind. Sci Am 281:112–117
12. Panksepp J (2003) At the interface of the affective, behavioral, and cognitive neurosciences: decoding the emotional feelings of the brain. Brain Cogn 52:4–14
13. Gallagher S (2000) Philosophical conception of the self: implications for cognitive science. Trends Cogn Sci 4:14–21
14. Gallagher HL, Frith CD (2003) Functional imaging of 'theory of mind'. Trends Cogn Sci 7:77–83
15. Damasio A (2003) Feelings of emotion and the self. Ann N Y Acad Sci 1001:253–261
16. Northoff G, Panksepp J (2008) The trans-species concept of self and the subcortical-cortical midline system. Trends Cogn Sci 12:259–264
17. Northoff G (2004) Philosophy of the brain. John Benjamins Publishing, Amsterdam
18. Roskies A (2002) Neuroethics for the new millenium. Neuron 35:21–23