Chapter 2

The brain as the subject's heir?

Overview

Chapter 2 critiques the claims according to which subjectivity is to be regarded as a construct or epiphenomenon of neuronal processes and thus one's experience of agency and freedom of choice should be seen as an illusion. First it is shown that the subjectivity of "experiential facts" cannot be reduced to objective or physical facts about brain processes. Likewise, the reduction of the intentionality of consciousness to relations of representation is refuted (2.1). Moreover, the identification of the subject with the brain leads to fundamental category mistakes which will be examined as the "mereological fallacy" and the "localization fallacy" (2.2). On this basis, a critique of the thesis of the powerlessness of the subject is developed (2.3). Finally, the summary analyzes the basic "naturalistic fallacy" of an objectifying account of consciousness which believes it can remove itself from its rootedness in the lifeworld (2.4).

Just as the world we experience, the experiencing and acting subject also becomes a product of brain processes from a reductionist standpoint. If the physical world is deemed actual reality, then the subject can, of course, only be allotted an epiphenomenal status. A rising choral song of materialist neurophilosophy heralds the message that our subjective experience is nothing else but the colored user interface of a neurocomputer, which even creates the illusion of the user itself.[1] Our experience of being the authors of our thoughts and actions is only part of this "grand illusion." Actual reality consists of the computational processes of the neuronal machinery running in the background:

> Our thoughts and our dreams, our memories and our experiences all arise from this strange neural material. Who we are is found within its intricate firing patterns of electrochemical impulses. (Eagleman 2015, 5)

[1] See Slaby (2011). Daniel Dennett was probably the first to claim that consciousness is "the brain's user illusion of itself" (Dennett 1991).

> The neurobiology of consciousness faces two problems: the problem of how the movie-in-the-brain is generated, and the problem of how the brain also generates the sense that there is an owner and observer for that movie. In effect, the second problem is that of generating the *appearance* of an owner and observer of the movie *within the movie*. (Damasio 1999a, 11)

The complex, even paradoxical structure of human self-consciousness, which is most difficult to grasp philosophically, is here subsumed with a flick of the wrist under the general neuroconstructivist thesis. If consciousness is only an "out-of-the-brain illusion," why not also the subject who has this conscious experience? One just needs to add a "meta-representation" to the inner representations (thoughts about thoughts, images of images), and self-consciousness is explained: we are only dream subjects within a dream. The brain is a "world simulator" and, at the same time, a "self-simulator." In Metzinger's "self-model" theory, the subject is consequently conceived as analogous to a pilot who believes that he experiences reality, while he is in fact placed into a flight simulator—and who is indeed himself only a product of this simulator, or a virtual self-model:

> The human brain can be compared to a modern flight simulator in several respects. Like a flight simulator, it constructs and continuously updates an internal model of external reality […]. However, there is a difference. […] there is no user, no pilot who controls it. The brain is like a *total flight simulator*, a self-modeling airplane that, rather than being flown by a pilot, generates a complex internal image of itself within its own internal flight simulator […]. Operating under the condition of a naive-realistic self-misunderstanding, the system interprets the control element in this image as a non-physical object: the "pilot" is born into a virtual reality with no opportunity to discover this fact. (Metzinger 2009, 107–108)

Of course, this comparison suggests the Cartesian picture of a pilot steering the body-plane—only to then refute this picture as a naive or dualistic self-deception. However, no serious philosopher nowadays claims that the subject or self should be regarded as some type of "non-physical object," a *thing* or an *entity* that could be distinguished from the person as a whole. And to be fair, even Descartes himself explicitly declared "that I am not lodged in my body as a pilot in a vessel, but that I am very closely united to it" (Descartes 1993, 93). Only if one would assume "the Ego" or "the Self" (writ large, as it were) to be a pilot or homunculus somewhere within the body, would it indeed be justified to speak of a "self-misunderstanding." But why should it be an error or an illusion if the airplane-system—or rather, the whole living being or the embodied person—is simply aware of itself *as itself*? There is nothing self-contradictory or illusionary involved, nor a "myth of the self" which Metzinger has avowedly set out "to shatter" (2009, 1). Hence, if he boldly claims

> that to the best of our current knowledge there is no thing, no indivisible entity that is *us*, neither in the brain, nor in some metaphysical realm beyond this world (2009, 1)

then the simple reply is that Metzinger searches for "us" in the wrong places. The indivisible entity that we are, is indeed neither in the brain nor in an otherworld, but it is quite visibly *our bodily being*—a living, functionally indivisible and self-aware organism. We do not exist a second time within our bodies or somewhere else. The "myth of the self" is just that: a myth. Like many neuroscientists and neurophilosophers, however, Metzinger prefers to keep on fighting against this "ghost in the machine" (Ryle 1949), for this lends more clout to his thesis of the "grand illusion":

> We are Ego-Machines, but we do not have selves. We cannot leave the Ego Tunnel because there is nobody who could leave. […] Ultimately, subjective experience is a biological data format, a highly specific mode of presenting information about the world, and the Ego is merely a complex physical event—an activation pattern in your central nervous system. (Metzinger 2009, 208)

Of course, the question immediately arises how Thomas Metzinger could become aware of living in the Ego Tunnel, if there is no escape from it—indeed, if there is even "nobody who could leave." A dreamer who becomes aware of dreaming can no longer be *only* a dream (this was already Descartes's bastion against an assumed "malign genius" who could deceive me of everything—except that I am the one who doubts). But be that as it may, let us return to the question of the self: granted, we may not "*have*" a self, but why should we not *be ourselves*, only because our self-awareness as living beings requires, as one necessary condition, an integrating activity of the brain? Later in his text, Metzinger himself concedes that it might also be possible to term the *organism*—as a self-organizing and self-sustaining system—a "self." In this case, the self, as he continues, would not be "a thing but a process":

> As long as the life process—the ongoing process of self-stabilisation and self-sustainment—is reflected in a conscious Ego Tunnel we are indeed selves. Or rather, we are "selfing" organisms: At the very moment we wake up in the morning […], [a] new chain of conscious events begins; once again, on a higher level of complexity, the life process *comes to itself*. (2009, 208)

From the point of view of embodied subjectivity which I will develop in this book, this seems a fairly acceptable position—provided only that we replace the provocative catchword "Ego Tunnel" by the more appropriate term "self-awareness." As I will argue further below, there is indeed an inherent continuity of life and awareness, or *Leben* and *Erleben.* Hence, in self-awareness, the life process of the organism in fact comes to itself, for it has always been a *self*-organizing process. But Metzinger does not seem really satisfied with this option. After all, he has already stated in his introduction that

> No such things as selves exist in the world. A biological organism, as such, is not a self. (2009, 8)

Therefore, he is now eager to assure that:

> True, upon your awakening from deep sleep, the conscious experience of selfhood emerges. […] But there is no one doing the waking up, no one behind the scenes pushing the Reboot button, no transcendental technician of subjectivity. […] Strictly speaking, there is no essence within us that stays the same across time, nothing that could not in principle be divided into parts, no substantial self that could exist independently of the body […] We must face this fact: We are *self-less* Ego-Machines. (2009, 208)

As we can easily see, in order to push the unwanted option aside, Metzinger needs to revive the Cartesian strawman once more: the reader should believe that in order to speak of a self, it must be a "transcendental technician" that steers the life process just as a user operates his computer. And instead of assuming the continuity of the living organism as the basic continuity of ourselves, there should be an "essence within us" which stays the same across time.[2] But this indivisible and bodiless Cartesian entity sadly does not exist (we must face it …), and Metzinger is glad to renew his illusion thesis.

Our first exposition of the reductionist claim has already pointed out one of its central weaknesses: the imputation of a Cartesian "Self" which no one actually supports, and the corresponding lack of a concept of the living being. Nevertheless, the concepts of an epiphenomenal or illusory subjectivity will now be critiqued in more detail in three steps. First, it will be demonstrated that subjectivity and intentionality cannot be reduced to physical descriptions of brain processes. The second step will examine the false conclusions and aporias to which the identification of the subject with the brain necessarily leads. In the third step, the claim of the subject's ineffectiveness and impotence will be refuted.

2.1 First criticism: the irreducibility of subjectivity

2.1.1 Phenomenal consciousness

The notion of a self-model implies that subjectivity or phenomenal consciousness is only an image or representation of the neuronal processes constructing it. However, the catchy term "model" only conceals the crucial problem: how could a physically implemented structure possibly give rise to consciousness of the world and of itself? After all, it is the same case with consciousness as with color: without our experience of it, science would not have any reason to even suspect its existence. To put it more pointedly: in a purely physically described

[2] On the continuity of the *embodied* self, see my publication entitled "Self across time: The diachronic unity of bodily existence" (Fuchs 2017a).

world, however complex its processes, something like consciousness simply does not show up, just like colors. In contrast to the brain, consciousness is not an object in the world—on the contrary, it is the presence of the world for a subject.

In his famous essay: "What is it like to be a bat?," Thomas Nagel has defended the resistance of subjective experience against its complete objectivization: even if we could fully describe the processes and behavior of a bat neurophysiologically, we would not have the slightest idea what it experiences or how it feels pain or ultrasound, in other words, what "it feels like to be a bat" (Nagel 1974). Therefore, there is basically, according to Nagel, an epistemological boundary for the neurosciences: subjective or experiential facts which are each only accessible from a unique perspective cannot be transferred completely into objective facts which can be observed by various individuals. The subject is the center of a world, and such centers cannot be found in a purely physical world, including neuronal processes.

It has become common to express this contrast in terms of the phenomenal or *first-person perspective* and the naturalistic, objectifying or *third-person perspective*. However, the source of the notion of perspective from an optical point of view should not allow us to forget that, in the case of the first-person perspective, more is at stake than just a particular angle, namely precisely "how it is" or "what it feels like" to be in a certain mental state, that is, an *elementary affective self-experience* before any self-reflection.[3] Subjectivity in this basic sense does not mean a perspectival view on contents or objects, connected to a conscious ego-experience; we are rather dealing with a primary bodily-affective self-feeling as the core of all conscious processes. Even before every perspective and cognition, there is a form of immediate, pre-reflective self-presence, an affectively colored familiarity of consciousness with itself, which may, according to Michel Henry (1963), also be designated "auto-affection."[4]

This self-affection may be taken to ground the *first-personal givenness* of every experience, which Zahavi (1999, 2005) has elaborated. Thus, any sensation, any

[3] For this, the term "what-is-it-likeness" has also come to use since Nagel's argument.

[4] This is in contrast to common higher-order or representational theories of consciousness (e.g., Carruthers 2005, Rosenthal 2005); on their critique, see mainly Zahavi (1999). In this respect, the analyses of Michel Henry are also comparable with concepts of the "Heidelberg School" (Henrichs 1970, Frank 1986, 1991) who assume a pre-reflective self-familiarity of consciousness as the basis of all higher-order reflective self-recognition. "Familiarity" (*Vertrautheit*) implies an affective element which is not explicitly thematized by Henrich and Frank, however. For an overview on phenomenological accounts of pre-reflective self-consciousness, see also Thompson and Zahavi (2007).

perception or action directed towards an object implies a tacit self-awareness without requiring introspection; it is given immediately, non-inferentially as mine:

> This first-personal experiential givenness is manifest in the very having of the experience. It is a givenness that obtains even when we are not explicitly aware of it […]. A conscious mental state is not merely conscious of something, its object; it is simultaneously self-disclosing or self-revealing. (Zahavi 2017, 198)[5]

Further, the basic affective self-awareness grounds the existence of *subjective or experiential facts*—for example, the fact that *I* experience pain, feel hunger, am happy or sad.[6] Thus, it is also the basis for everything existentially meaningful, for what constitutes my personal concerns and cannot be replaced by taking a general or scientific point of view.[7] Can such subjective facts be reduced to objective ones, for example, facts which can be described in neurobiological terms? Is it possible to describe the fact that I am now feeling pain as a certain neuronal activity pattern without its losing its significance? No, because even the seemingly unproblematic re-formulation "Thomas Fuchs feels pain at this moment" no longer expresses the fact that it is *my* pain and that it is I *myself* who suffers from it.[8] Even if this statement from the third-person perspective were reliably true in all cases (e.g., on the basis of the simultaneous observation of my brain processes), it lacks the decisive feature of subjectivity, namely that *I myself* am that T.F., about whom this statement is made. This would be all the more true for an exact description of the physical processes in the brain of T.F.—nowhere in it could the *mineness* of the pain be found. Between both manners of stating this, there is an ontological leap. The reality of my pain is of *a basically*

[5] Zahavi's concept of pre-reflective self-awareness, also termed "minimal self," does not emphasize its affective aspect which is highlighted in my account, yet certainly does not exclude it.

[6] These are not primarily *propositional* facts, that is, facts that are expressed in propositional terms ("I feel pain," "I am sad"). The feeling of pain or sadness is before any verbalization, in which I could also be mistaken (e.g., because I feel not precisely "sadness" but some related emotion, say, disappointment). Primary experiential facts, as such, are "immune to error through misidentification" (Shoemaker 1968).

[7] "We feel that even if all possible scientific questions be answered, the problems of life have still not been touched at all," writes Wittgenstein (Tractatus Logico-Philosophicus, 6.52; Wittgenstein 1961).

[8] Here I draw from Hermann Schmitz's analysis of subjective facts: "A fact […] is *subjective*, if at most *one*, and only on his own behalf, can state it, while others may well speak about it with unequivocal labelling, but never ever can state what is meant" (Schmitz 1995, 6; own translation).

different kind to the reality of objective physiological facts—and nevertheless it is no less "real" than these.

Facts of self-experience cannot be transferred into objective facts without a decisive loss. And indeed not so much because of their special qualities or "qualia,"[9] but rather, above all, because of their subjectivity itself: it constitutes an *absolute epistemic asymmetry of facts*. The natural scientific reduction is based, as already presented in the "Introduction," on stripping subjectivity from the experienced facts and reducing the remainder to elementary physical processes. It thus transforms what is subjectively experienced into objective statements, which is connected with a loss and alienation, but which is practical and successful for the purposes of explaining and predicting nature. The reduction fails, however, when subjectivity as such is at stake. Even if it could be proved that subjective experience is always produced by certain neuronal activities (we will see whether this is in fact the case), the explanation would still remain incomplete—the radically new ontological characteristic of the subjective itself could only be accepted and not be explained further from the physical processes.

The principal asymmetry between subjective and objective facts also manifests itself in the performative function of certain speech acts. The statement "I promise to visit you tomorrow" is obviously not equivalent to the statement "Somebody promises to visit you tomorrow, and the person who promises that is Thomas Fuchs." The act of promising as a performative self-commitment can only be expressed in the first person; the report about the promise of a third person, even if it is completely correct, does not include this commitment.[10] It becomes clear that the I-statement of a speaker cannot be transformed into the report about a third person without crucial semantic loss. For the fact that the promise concerns *me* and my affectively colored experience of self-congruency and self-commitment, which I put in the balance, is eliminated from the objectifying description. The performative effect of certain speech acts thus marks a subject as the irreducible center of self-related meanings and of

[9] The problem of the "explanatory gap" (Levine 1983) in philosophy of mind is usually explained on the basis of "qualia": even if we were certain that phenomenally conscious states are identical with brain processes, we would not have a scientific explanation for the fact that these processes are experienced in the special qualitative way of pain, color, sadness, and the like. However, the qualia problem only concerns a partial aspect of subjectivity; in my view, it does not constitute the decisive explanatory gap which is rather based on the "mineness" of any conscious awareness as such.

[10] On this, see Ricoeur's analysis of the subjectivity of performative speech acts (Ricoeur 1992, 42–43).

being affectively concerned. In other words: based on the absolute epistemic asymmetry of *facts*, there is also an absolute performative asymmetry of *certain actions*.

2.1.2 **Intentionality**

Whereas our line of argument at first applied to subjective experience, as it manifests itself in conditions of pain, hunger, sadness, or the like, our last considerations went beyond that. Subjectivity is not merely state-like, it is moreover essentially oriented *to what it is not itself*: it is open to the world, related to objects, and directed to contents and meanings. Experiential states which are of such types that they are directed to something, that is, perceptions, thoughts, wishes, ideas, or memories, possess the characteristic of *intentionality*. That is to say, they have an intrinsic content to which they relate and which can be expressed by a that-clause (e.g., promising "that I will come tomorrow"; believing "that Monica is wrong"; wishing "that the rain stops"; etc.). In other words, intentionality opens the dimension of *sense and meaning*.

It is obvious that the intentionality of consciousness represents a serious problem for a physicalist reduction—more serious, in fact, than its subjectivity—because experiences with missing or weak intentional content, such as pain or moods, could still be objectified as "mental states" and thus possibly be equated with neuronal processes. Intentionality can, however, no longer be adequately defined as a mere mental state; for what is meant or intended by them belongs to the definition of intentional acts. The mental state of the intention to buy a book does not exist independently of the book, the way to the bookshop, the purchasing process, and so on. In other words, it presupposes its *embedding in a situational and meaningful context.*[11] A definition of intentional acts, independent of object and context, would, however, be the precondition for its description as states of the brain. Physical processes, such as the activations of neurons, can, as such, not be aimed at a context, and the imaging of brain activities during intentional acts cannot basically capture the direction of their sense.

2.1.2.1 Intentionality and phenomenal consciousness

Nonetheless, in the analytical philosophy of mind, the naturalization of intentionality is attempted in two steps. First, the phenomenal characteristics of

[11] On this, see also Searle: "semantic contents, that is, meanings, cannot be entirely in our head, because what is in our heads is insufficient to determine how language relates to reality. […] If the meaning of the sentence 'Water is wet' cannot be explained in terms of what is inside the head of speakers of English, then the belief that water is wet is not a matter solely of what is in their heads either" (Searle 1992, 49).

consciousness and subjectivity, as so-called qualia, must be separated from the intentional characteristics. Intentional meanings, as Chalmers (1996) argues, for example, could then be construed in terms of a functionalist theory which would explain them as neuronal representations. Certain neuronal system states are, through the previous history of the brain, functionally connected with certain configurations of the environment. That is why, in each case, they produce the suitable output for a certain input and make the intentionality of consciousness dispensable. Functionalism thus seems the suitable strategy for reducing intentionality. The qualia problem would then be left as the only "hard problem of consciousness" (Chalmers 1996)—but this problem could as well be ignored as negligible for the overall course of the world. A functional definition, for example, of pain would consist of the connection of physical input (tissue damage or trauma) and behavioral output (aversive or avoidant behavior). The *feeling* of pain is irrelevant for this connection.

It has already been shown that, with the problem of subjectivity, there is indeed more at stake than certain individual qualities, such as "red" or "warm," that is to say, subjective experience as such. Is it possible to separate intentionality from subjectivity? Is the *experience* of meanings in principle a dispensable addition? The claimed separability of subjectivity from meaning presupposes a reductionist re-definition: "meaning" would then consist only of the two-place assignment of sign and signified, or *representatum* (the representing internal state) and *representandum* (the represented part of reality), and this assignment would be purely functionally realized by the regular connection between the input and appropriate output of the brain. However, Galen Strawson has emphasized that meanings only exist *for someone:* "Meaning is always a matter of something meaning something to someone. In this sense, nothing means anything in an experienceless world. There is no possible meaning, hence no possible intention, hence no possible intentionality, on an experienceless planet" (Strawson 1994, 208–209).

Intentionality is thus a three-place relation: *something* means *something for somebody.* "I believe that Monica will come" puts (1) Monica in relation to (2) an act of supposition, which can only be attributed to me (3) as a conscious person. Intentional acts and attitudes are something whose meaning is *experienced* and which, thus, necessarily belongs to a phenomenal consciousness. Wishing or wanting something, remembering or recognizing something, understanding words—all these possess a certain quality of "what it is like" to experience this state. Seeing an apple is different to imagining an apple (Zahavi 2003). Each is connected with a particular way of experiencing and self-experiencing—just like experiencing pain, hunger, or sadness. Thus, intentionality and subjectivity cannot be separated from one another.

2.1.2.2 Intentionality and representation

The decisive notion, which is, nevertheless, intended to achieve the naturalization of intentionality, is the concept of *representation.* Let us therefore consider this central concept of cognitive neurosciences or neuroinformatics more closely.[12] *Neuronal representations* should depict an external fact or a set of facts in a neuronal system in such a way that they can represent ("mean") this in the cognitive operations of the system. All information about the fact is mirrored in representing patterns of neuronal activity and can, as such, be further processed. They are usually regarded as the basis of "*mental representations*"—the contents of consciousness. Renewed pictures of the neuronal representations on a higher level, in other words, meta-representations, would then be the basis for reflective processes. Thus, the intentional contents of consciousness would be physically realized and as such could have effects on the output of the system, that is, on the behavior, without the phenomenal intentionality of a subject being required for that.

Searle has shown that in reality only an "as-if" intentionality is constructed in this functionalist account (Searle 1992, 78–84). For a meaningful connection cannot be ascribed to functional, rule-consistent procedures without there being someone who *understands* this connection. In order to illustrate this, Searle has developed the thought experiment of the "Chinese Room," which has become commonly known but shall nevertheless be briefly described here (Searle 1980):

> Imagine that someone who does not understand a word of Chinese is locked in a room, in which there is a program with all the rules for answering questions in Chinese. The person now receives questions that are passed into the room which are written in Chinese symbols ("input" into the system) and works out completely correct answers with the help of the program, which he then returns ("output" of the system)—of course, these are purely rule-consistent and he does not understand any of them. Let us presume that the program is so perfect and the answers are so good that even a Chinese person outside the room would not notice the deception. Nevertheless, one could not say about the man in the room that he *understands Chinese.* The semantic content or meaning of the language thus contains more than its mere grammar and syntax.

Searle's "Chinese Room" is, of course, the image for an information processing machine in which a central processor works according to the algorithms

[12] Main proponents of representationalism in philosophy of mind are, for example, Dretske (1995), Tye (1995), Lycan (1987), and Metzinger (2003). However, the concept is common in most neurocognitive theories as well as in accounts of empirical studies.

of a program ("If you get input X in the context Y, then give output Z"). The machine functions completely adequately as a system, but, nevertheless, it lacks the decisive characteristic of intentionality, namely the semantic content—*experienced meaning* or *comprehension*. Hence, our understanding cannot be reduced to program procedures or information processing in the brain.

This can be transferred to all technical cybernetic systems: a torpedo is programmed so as to detect a moving target and pursue it. We can also say that the object is "represented" in its steering system. However, this representational function only exists *for us*, namely on the basis of our previous construction and programming, which places the torpedo in a regular connection with a target object. The steering mechanism allows the torpedo to make corrections in movement, by means of which it finally reaches its target. Nonetheless, it would, of course, be nonsensical to say that the torpedo "seeks its target," that is, in fact, it has an intentional and time-spanning relation to its target object. Every correction only serves the internal set-point regulation of the mechanism and occurs purely momentarily without relating in any way to a target anticipated *as such*. For this goal itself, the mechanism remains blind and deaf. If it reaches it, the program is simply over—its purpose is, however, only "fulfilled" from our point of view.[13] The "representation" of external facts in a system is, thus, completely different to the intentional directedness to these facts.

The notion of representation is meant to eliminate this experienced significance—that is why it is so cherished in neurophilosophy. In fact, however, it is *only we ourselves* who can ascertain the representation of one fact or event by another fact; it does not exist *as such*. As a rule, contexts of representation are created by us. The map of a country which we produce represents a landscape; a portrait, a human being; and a sentence, a set of facts. In an improper sense, representations may also be ascribed to objects of nature as the result of causal connections—in this sense, a track in the snow "represents" an animal, smoke "represents" fire, and the rings in a tree trunk's cross-section "represent" the life years of the tree.[14] In all these cases, however, the representation exists *only for us* who can establish the context of meaning, insofar as we dispose of intentionality. For nothing prevents us from attributing representations not only to the smoke or the growth rings but to all effects traced back to a certain cause: the warmth of the earth at night "represents" the daily solar radiation, the

[13] On this crucial difference, see also Jonas's critique of "cybernetics and purpose" (Jonas 2001, 108–127).

[14] Both Dretske (1995) and Tye (1995) take the growth rings as an example of a "natural" representational relationship on the grounds that the number of rings causally co-varies with the number of years. For a poignant critique, see also Bennett and Hacker (2003, 142).

tides "represent" the moon's gravitation, and the stomach mucosa "represents" the incoming food by producing a regular output, namely gastric acid. So if representations existed "as such," in the subjectless nature, it is obvious that they would exist everywhere as well as nowhere.

Each semiotic relation is three-place too: *something presents a sign of something for somebody*.[15] That is why in a computer *as such* nothing more takes place than transitions from one electrical state to another. Only the programmer or user can interpret these processes as symbol manipulations or information processing, thus *lending* them *meaning*. Briefly: in a world without subjective experience there are no longer signs, nor symbols or information, representations or meta-representations, meaning or sense. "Reading" representations "into" a purely objective causal connection of natural processes is, in this respect, a conceptually unsound manner of speaking, intended to give the neuronal processes an appearance of intentionality.

One can, admittedly, attempt to define representation in terms of a three-place relation *without* a subject, as Metzinger does:

> Mental representation is a process whose function *for* the system consists in representing actual physical reality […] [An] internal state *X* represents a part of the world *Y for* system *S*. (Metzinger 2003, 26)

Certain neuronal processes, as representata, thus depict an external state for the system, by which Metzinger means an information processing system such as a human organism or its brain (2003, 24–25). But this seemingly three-place relation cannot be maintained. The preposition "for" indicates either the reference to an intended goal or purpose ("what is this good *for*?") or to a subjective point of view ("*for* me it is clear that …"). Both kinds of relation cannot apply here, for a subjectless system neither pursues goals (like the torpedo, it only passes through regulations and adaptations, but is indifferent to its state), nor does it have a point of view. A goal could only be ascribed to it by its engineer or designer, but this external view would not solve the problem. Nonetheless, Metzinger speaks of the "for" relation as a "teleological criterion" and regards mental representations "as internal tools, which are currently used by certain systems in order to achieve certain goals" (2003, 26–27). Granted, at present these can only be biological systems:

> Artificial systems—as we knew them in the last century—do no possess any interests. Their internal states do not fulfil a function for the system itself, but only for the larger unit of the man-machine system. (2003, 27)

[15] Peirce's definition of the sign is in accordance with this: "A sign, or representamen, is something which stands to somebody for something in some respect or capacity" (Peirce 1932, 228).

However, neither an artificial nor a biological system, *taken only as a cybernetic system*, has an "interest" in "achieving certain goals." Granted, it may be in the position to fulfill certain functions—be it for human purposes or for its own preservation. But these functions may only be ascertained from the outside. As long as nothing is *at stake for* the system and it does not have *concerns* or goals, his functionality does not imply any teleology. It is not "too cold" in the room for a thermostat, nor for a brain, and a torpedo does not "experience failure" when it misses the ship.

In contrast to machines, a biological system admittedly perishes if its "representations" are not functionally adequate. They have thus a function for the preservation of the system—a function which may be traced back to a causal history of evolutionary selection.[16] However, one can still not talk about interests and goals which the biological system pursues, rather only about a natural causal history which produced systems of a kind that their internal processes may be described *from our point of view* as "functional" in the sense of self- or species-preservation. *For the systems as such*, it does not matter at all whether they perish or not (of course, as long as they do not have *subjectivity*, and thus concerns and interests—but this is not implied in Metzinger's definition). With this, however, the precondition for a three-place concept of representation, which could refer to a subjectless system, is lost. Metzinger's definition can then imply no more than that the neural system produces certain activation patterns or "data formats" which *we* can interpret as "representations" and as tools for self-sustainment. Whichever way you look at it, the representational relation—something *stands for, points towards*, or *means* something else—cannot be reinterpreted as a functional–causal connection, without there being subjects *for whom this is functional.*

A neuroscientist may nevertheless continue to speak about "representations" or "maps" in the brain in the sense that certain neuronal activation patterns are causally connected and correlated with a perceived object, an imagined object, or the like. He may also use such observed correlations to make inferences about the present perception or imagination of the owner of the brain. However, these patterns *as such* are not therefore *symbols* of objects, *they do not refer to them, do not mean them, and do not represent them*—no more than a tree presents its years of age in its

[16] This is the strategy of "teleofunctionalism," to which also Metzinger consents (2003, 27); on this, see Block (1978), Lycan (1987), and Millikan (1984). According to Millikan, the project of teleofunctionalism is to derive functions (and accordingly malfunctions or misrepresentations) from a causal natural history, or in her own words, "to let Darwinian natural purposes set the standards against which failures, untruths, incorrectness, etc., are measured" (Millikan 1991, 151). The concept and the critique it has received cannot be dealt here in more detail.

growth rings. There are no representations of the outer world in the brain, either in the semantic or the iconic sense of the word.

Should one not at least speak of traces of *memory* as representations of what is experienced in the brain? Without them, the person could surely not remember their knowledge of, say, World War I. Well, remembering something realizes an *ability*, such as the ability to recite a poem or to play a sonata by Schubert on the piano. When learning a poem, the brain undoubtedly develops the preconditions for a person being able to remember it later, for example, certain synaptic connections and dispositions for neuronal excitation. The poem is, however, not "stored" in the brain as a "representation," no more than their memory of the dates of World War I or of their voyage to Morocco, for the brain contains neither sentences nor pictures. Sentences in books represent facts *for us*, pictures in photo albums represent memories *for us*. However, there is no homunculus in the brain who would be able to grasp neuronal patterns of activity *as* representations, to see them *as* pictures or to read them as traces of memory. Neither rings in the tree, nor tracks in the snow, nor neuronal activity patterns in the brain are, *as such*, "representations" of past events.[17]

Hence, a valid concept of representation in the cognitive neurosciences would have to include the point of view of the observer. Representative connections can only be ascertained from the perspective of researcher subjects, who are, in addition, dependent on the statements of their test subjects in the first-person perspective, if they wish to arrive at correlations with subjective experiences. Talk about functions or functional connections is, for its part, *necessarily teleological*: in order to be able to determine the function of certain processes within

[17] Again, see Bennett and Hacker (2003, 154–171). Similarly Edelman and Tononi reject a representationalist account of memory: "Representation implies symbolic activity, an activity that is certainly at the center of our syntactic and semantical language skills. It is no wonder that in thinking about how the brain can repeat a performance—that it can, for example, call up what may appear to be an image already experienced—we are tempted to say that the brain represents. The flaws in yielding to this temptation, however, are obvious: There is no precoded message in the signal, no structures capable of the high precision storage of a code, no judge in nature to provide decisions on alternative patterns, and no homunculus in the head to read a message. For these reasons, memory in the brain cannot be representational in the same way as it is in our devices" (Edelman & Tononi 2000, 94). Instead, memory should be regarded as a "system property," which enables the brain to dynamically react to current situations and, on the basis of established neuronal dispositions, to activate varying response patterns not in a replicative, but in a creative way. In short, memory is never based on fixed "engrams," "copies," or "representations," but always recreates *similar* images or actions.

a system, I must, as an observer, presuppose the sustainment of the system as a purpose. Hence, if the notion of representation should serve to eliminate subjective experience or to identify subjective with brain states, the neuroscientist loses sight of the prerequisite for his research: his own subjectivity. However, since the neurocognitive notion of representation may hardly be purged from its semblance of objective givenness any more, it seems more reasonable to replace it generally, for example, by the term *pattern* and *pattern resonance* (on this, see section 4.2)

Let us sum up: ascribing intentionality to certain (not all) processes of consciousness identifies its inherent directedness to objects. However, intentionality cannot exist without subjectivity. Although the *performance* of intentional acts is linked with certain organic processes of a living being, its content, namely, "grasping something *as* something," does not tally with any physical or physiological description. There is, in fact, no meaning, no sense without subjects. The concept of representation is intended to indicate a two- or three-place relation, which could be described purely functionally. Nevertheless, each relationship of representation only exists for a person, who recognizes and interprets it as such. A picture is not a picture without someone who grasps it *as* a picture; a sign means nothing unless there is someone who understands it as a sign; a track refers to nothing without a tracker: the concept of representation cannot replace subject-dependent intentionality.

2.2 Second criticism: category mistakes

2.2.1 The mereological fallacy

Let us now examine the category mistakes and fallacies which result from the identification of the subject with the brain. These include, first and foremost, the neuroscientific practice of personalizing the brain and ascribing to it the most varied human activities. Brains can then, for example, "recognize faces" (Caharel et al. 2009), "perceive taste with all senses,"[18] but also "perceive alcohol" (Hodge et al. 2006). The inferotemporal cortex "identifies objects,"[19] the brain "decides when to work and when to rest" (Meyniel et al. 2014), and it even "recognizes itself as the subject of recognition" (Northoff 2004a, 17). If one reads neuroscientific literature, one can almost come to the conclusion that the brain genuinely calculates, believes, interprets, construes hypotheses,

[18] Science Daily 2016 (https://www.sciencedaily.com/releases/2016/08/160831133706.htm).

[19] MIT News 2015 (http://news.mit.edu/2015/how-brain-recognises-objects-1005).

recognizes, and decides. The category mistake occurs so often that Bennett and Hacker (2003) have given it a name of its own, namely, that of the "mereological fallacy."[20] A part of the organism, the brain, or one of its subsystems, thus has psychological and personal activities ascribed to it, which, in fact, only belong to the person as a whole. Examples abound, but I give only one more of them here:

> This simple fact makes it clear that you are your brain. The neurons interconnecting in its vast network, discharging in certain patterns modulated by certain chemicals, controlled by thousands of feedback networks—that is you. And in order to be you, all of those systems have to work properly. (Gazzaniga 2005, 31)

Well, of course I am not my brain—for my brain is certainly not married, not a psychiatrist, and it has no children. Even worse, it does not see nor hear anything, it cannot read or write, it cannot dance or play the piano, and so on. Thus, I am rather glad not to *be* my brain, but to only *have* it.

However, the personalizing language is not only meant figuratively or metaphorically, as the defense of this position is often articulated—on the contrary, it is precisely a successful naturalization which requires infiltrating intentional vocabulary into the description of subpersonal processes. For what could be explained about man if one only described monotonous, electrochemical processes on his neuronal membranes? The dissection of the live whole into micro-processes must, at least verbally, be undone, in order to reach the level of perceptions, motives, and actions again. The neurosciences, for that reason, attempt to insert a "hybrid" level in between which blends the physical and intentional descriptions, thus, to a certain extent, implanting personality in the brain.

That seems less problematic the more one goes over from actions to "pure" cognitions. Does the brain write? Does it hear, does it see?—Hardly. But does it think perhaps?—That may well seem so. Nevertheless, what could we make of a sentence such as this: "Peter's brain intensely deliberated about what it should do. When it could not find a solution, it decided first of all to wait and see." If thinking, feeling, and deciding were, in fact, activities of the brain, this would not be a ridiculous sentence, rather a quite meaningful one. Yet we rightly ascribe such activities to Peter, and not to his brain, because they are simply not "cognitions" or "mental states" in which Peter is, rather they are *life acts* which can only be ascribed to Peter as an embodied and conscious being. Reflecting, feeling, wanting, and deciding—none of these can be found at the physiological

[20] Mereology means the relation of parts and whole (from the Greek *méros* = part).

level of description *because these concepts do not exist there at all.* It is not wrong for empirical reasons to speak of the thinking, feeling, or perceiving brain—it is much rather conceptual non-sense. Erwin Straus formulated this insight briefly and appropriately: "Man thinks, not the brain" (Straus 1956, 112).

In their critique of the mereological fallacy, Bennett and Hacker (2003, 71–72) show that, behind the "as-if" subjectivity of the brain, there is again a latent Cartesianism: the "I" or "Ego" is thought of by neuroscientists as a substantialized, supposedly autonomous, freely acting center of decision, which is then declared to be non-existent: "It is not the Ego, but my brain, which has decided." This still assumes that there could be something like a Cartesian "Ego" making decisions. This Ego, the non-material soul is thus toppled and in its place comes the brain, only to immediately do the same as the Ego in Descartes, namely, to putatively imagine, to perceive, and to decide. Nevertheless, brains think or decide just as little as bodiless Egos—in both cases, one part is put in the place of the whole. This does not change if the Ego is replaced by "consciousness" or the "mind," as long as these concepts are, for their part, understood in the sense of a bodiless inner world. However, consciousness is a characteristic of living beings or, more precisely, an *enactment of life.* It manifests itself in life utterances and activities which are experienced by the living being as a whole and can be recognized by others in its behavior: being frightened, afraid, or happy, reflecting, speaking, writing a letter, or playing football.

That seems to be just a matter of course, which it is not, however. Even for John Searle, mental states are "simply higher-level features of the brain" and consciousness is "an emergent property of the brain" (Searle 1992, 14). On the other hand, shortly afterwards, he emphasizes that "the ontology of the mental is essentially a first-person ontology. Mental states are always somebody's mental states" (p. 20). But *somebody*, that is a person, therefore not "an Ego," "a consciousness," not to speak of a brain; it is rather a complete human being of flesh and blood. Can we, nevertheless, ascribe somebody's mental states to his brain? No, this is where Searle is contradictory: consciousness is a feature of human beings, that is, of organisms, not of brains. A neuroscientist may well be able to ascertain indications of a person being conscious in her brain—however, in order to find out whether she actually is conscious, he must observe her embodied behavior or engage in interaction with her. The brain may well be the central place for physiological processes which are necessary for her being conscious, but *it is not aware*, it does not perceive, it does not move, it does not get angry or feel happy—all of those are the activities of *living beings who are conscious.*

The basic problem of neurobiological research into consciousness consists, when all is said and done, in the *reification of consciousness itself*. It then no longer appears as an activity of living organisms, no longer as a relationship between subject and world which transcends the boundaries of the body. It is rather transferred into the objective world, as if it were an object in spatiotemporal reality which could be physically described or, at least, made indirectly visible by physical means. This leads us to a further fallacy.

2.2.2 **The localization fallacy**

A category mistake connected with the mereological fallacy consists in localizing single phenomena of experience in specific brain areas—we can speak of the "localization fallacy." According to it, visual perceptions are produced in the visual association cortex, fear in the amygdala, or memories in the temporal lobes. Constantly, new areas are found for all types of mental phenomena—pain, sadness, racist prejudice, deliberate deception, self-criticism, taking another's perspective, empathy, indeed even personality traits.[21] This research program is, first and foremost, based on imaging techniques which reflect the specific brain activities in vivo and seem to suggest that mental functions should be located in certain areas of the brain.

> The confrontation between localizational and holistic paradigms in brain physiology goes back as far as the eighteenth century. For a long time, localization theory was discredited by the "phrenology" of Franz Josef Gall (1758–1828), who speculatively related features of character, such as love of children, domesticity, or superstition, with certain areas of the cerebral cortex and corresponding protrusions of the skull. Albrecht Haller (1708–1777) and later Pierre Flourens (1794–1867) proposed a contrasting, holistic theory of the function of the brain, the so-called equipotential theory, according to which the complete brain always takes part in mental functions (Hagner 1997, 89–92, 248–50, Karenberg 2009). By means of the discovery of brain areas, whose failures are responsible for motor and sensory aphasias, Broca (1861) and Wernicke (1874), however, contributed greatly to the rehabilitation of the localization project, which enjoys particular success today. Accordingly, theories of the *modularity* of the mind (Fodor 1983, Pinker 1997), implying the construction of consciousness from separable single functions, are still preferred in cognitive science.

[21] See, for example, Phelps et al. 2000, Vogeley et al. 2001, Langleben et al. 2002, Etkin et al. 2004, Eisenberger et al. 2005, or Singer and Lamm 2009.

Undoubtedly, the localization theory has its own justification. The brain is regionally specialized; various neuronal areas and centers fulfill different functions. For this reason, it is also possible to connect certain *features or components* of conscious processes with local activities. Thus it is possible, by means of brain imaging and other procedures (single neuron recording, electroencephalography (EEG)), to ascertain with high probability whether someone is speaking silently to himself, imagining different categories of visual objects, adding or subtracting numbers, paying attention to a vertical or horizontal patterns of stripes, is preparing to press the right or the left button before him, and also whether a person is feeling pain, fear, or happiness (Edelman et al. 1998, Cox & Savoy 2003, Kamitani & Tong 2005, Soon et al. 2008, 2013). This is, however, only possible if corresponding correlations have been established by imaging beforehand, namely according to the information given by the test persons. Such advances are based on the functional specialization of the regions of the brain.

On the other hand, none of these regions is per se capable of producing the complex achievements of integration which are the basis of processes of consciousness. In fact, widely distributed brain areas and centers outside the cortex also contribute to this, so that a dynamically changing network of neuronal assemblies and activity patterns spread over the whole brain is involved in a special subjective experience.[22] Last but not least, the unsolved "binding problem"—the question of how the scattered activities and processing paths are reintegrated, as, for example, in unified intermodal perceptions (see 1.3.1)—points to the limitations of the localization paradigm (Uttal 2001).

According to the classic cognitivist or modular view, the brain implements encapsulated mechanisms for cognizing (perceiving, planning, evaluating, decision-making, etc.). Each module is believed to be responsible for computing an independent cognitive function, largely unaffected by the working of other modules and disconnected from bodily and environmental processes. This conception still fuels experimental cognitive research, not least because of its suitability for isolated study designs. However, it has now come under

[22] Edelman and Tononi (2000, 139–142) have proposed the "dynamic core hypothesis," according to which conscious states emerge from an ever-changing functional cluster of networks, characterized by strong interactions and "reentry" feedback mechanisms, and situated mainly within the thalamocortical system. "A dynamic core is therefore a process, not a thing or a place, and it is defined in terms of neural interactions, rather than in terms of specific neural locations, connectivity or activity" (2000, 144). What is neglected in this theory, however, is the role of body–brainstem interactions for the emergence of consciousness; this will be investigated in section 4.1.

growing criticism as being inadequate for the distributed functioning of the central nervous system, multitasking at every level and highly dependent on contextual variables (Van Orden et al. 2001, Hardcastle & Stewart 2002, Gibbs & Van Orden 2010). Therefore, the modular model is increasingly replaced by thinking in overarching functional systems and highly flexible brain connectivity patterns, where the same cortical or subcortical area may be co-opted into different functions depending on which of its interconnected networks is activated (Friston et al. 2003, Sporns et al. 2005; for an overview, see Cosmelli et al. 2007).

This also corresponds to the complexity of experience itself: all terms for special functions, such as seeing, hearing, thinking, feeling, wishing, and so on, single out components of consciousness, whereas factually subjective states of experience always remain holistic. Thus, all perceptions are not only embedded in a bodily background experience, but are also connected with feelings, memories, and linguistic concepts. There is no "pure" pain, no "plain" seeing or hearing. Conscious experience is not put together from components at all; it is, conversely, a *primary unified process* or a "*stream of consciousness*," which differentiates into specific activities and achievements according to the particular demands of the situation. Hence, brain functions may best be conceptualized along two polar organizational principles: functional segregation *and* functional integration. Their interplay is enabled by connectivity and distributed neuronal assemblies that transiently oscillate at the same frequency (Friston 1994, Cosmelli et al. 2007).

For that reason, however, talking about circumscribed "neuronal correlates of consciousness" is not appropriate. It implies that phenomena such as perceptions, feelings, or thinking processes could be isolated from the holistic activity of consciousness. These phenomena, however, are not states which can be isolated; they rather presuppose a *subject* that perceives, feels, thinks, and so forth. However, what kind of "correlate" subjectivity has, how far its organic base extends, and whether it does not include the complete organism, is still unexplained up to now. As long as this is not the case, the search for correlates of consciousness still remains at a speculative stage (Cosmelli et al. 2007).

> Noë and Thompson (2004) have pointed out that even in the best studied subsystem of the brain, namely the visual cortex areas V1–V5, it is not possible to unambiguously attribute visual content to certain neural assemblies. The reason is that even with regard to the same object, the activity of these neurons depends on the living being's body posture, behavior, state of attention, and the relevance of the object for current tasks, in short: on the overall state of the organism in relation to its environmental context.

> Moreover, each perception of a moving object contains not only the object itself, but also its motion dynamics, the background of the visual field, the eye, head, and body movements by which one follows the object, one's proprioceptive body awareness, and so on. Thus, perception is not a momentary snapshot of a stimulus configuration, but rather a dynamic, intentional, and attention-directed process which ultimately includes the whole system of brain, body, and environment. The search for neural correlates of consciousness can therefore only grasp certain partial components, not perception as a situated, bodily, and spatial process.

If attempts toward localization of consciousness or conscious functions lead to impasses, one may ask what misleads neuroscientists to localization fallacies time and again?—Above all, three kinds of observations contribute to this:

1. To begin with, it is specific *function failures* as a result of local lesions in the brain which seem to pinpoint the "seat" of the function in the relevant area. Because of the high plasticity of the brain, however, lost functions can in many cases be taken over by other brain areas. But even apart from that, the failure as a result of a lesion allows at best for the conclusion that the area represents the *necessary*, but not the *sufficient* condition for a function. There are always other areas and connections required within the complete neuronal system, as we have seen earlier in the case of perception. Hence, it is not functions that may be strictly localized, but only disturbances of functions.
2. The new *imaging techniques* seem to establish the place of the function in vivo. In a world of pictorial media, neuroscience has developed its public power of persuasion not least by means of its colorful staging. That is why it is all the more important to know the methodical limitations of these techniques.

   First, imaging techniques do not in any way measure neuronal activity as such, rather indirect parameters, as, for instance the BOLD signal (the blood oxygen level-dependent signal, i.e., increased blood flow and oxygen use in certain brain areas, from which the neuronal activation can be inferred) in functional magnetic resonance imaging (fMRI). In order to create a sufficient contrast, the basic activity of the brain is determined in advance and then "subtracted" so that the locally increased activations emerge. Thus we are not dealing with "images of the brain," rather of the visualizations of statistic calculations, that is, scientific constructs produced in an intricate manner. Further, mean values are formed from greater samples of test persons since no significant results can be individually gained as a result of the extremely limited differences in local activity. Not surprisingly,

the validity of the achieved correlations has also been strongly questioned, for example in affective neuroscience (Vul et al. 2009).

Moreover, it is not in any way clear whether the experiential phenomena investigated correspond to the most colorful flashing structures. In the case of pathological phenomena, local increases in activity can also correspond to secondary, compensatory reactions to the actual functional disturbances at another place. In any case, all other brain regions, in which nothing appears to happen in the image, are active at the same time and in various ways involved in the experience and the function. Thus, the resting state of the brain, a basic activity spread over the cortex and known as the "default mode" (Raichle et al. 2001), seems to represent the basis of a background experience, on which specific activities of consciousness can only develop. Finally, Anderson and Pessoa (2001), in a meta-analysis of 2603 fMRI studies in 11 task domains (e.g., vision, audition, attention, emotion, language, memory, action execution, etc.), found that in fact most regions of the brain are involved in supporting multiple tasks and can perform different operations under different circumstances, again pointing to the limits of the localization paradigm.

What the images actually show and what really happens in the brain thus require careful interpretation. Moreover, imaging occurs in laboratory situations, where the relation of conscious processes to the environmental context remains largely excluded, as does their prehistory and their temporal course. These aspects are, however, essential features of consciousness. The technique of imaging thus, as it were, freezes the stream of consciousness and isolates it from its context. If one takes all these methodical limitations together, data on the local metabolic activity of the brain can to a certain degree reflect its functional specialization, but it can only offer limited *indications* of ongoing mental processes.

3. The localizability of mental functions seems to be impressively shown by the fact that certain conscious phenomena can be evoked by direct electrical stimulation of the brain (see Selimbeyoglu & Parvizi 2010 for an overview). Thus, in the 1960s, the neurosurgeon Wilder Penfield succeeded in triggering, by means of targeted stimulation during brain surgery, the kind of experiences in conscious patients that are known as epileptic auras (Penfield & Perot 1963). Among these experiences were changes in perception (distortions of sounds or visual objects, experiencing déjà vu), feelings of pain, fear, sadness, or disgust, as well as memory flashbacks, voices of familiar persons, well-known melodies, or fragments of experienced scenes. Prior to the neuroimaging era, such brain stimulation experiments provided the most direct evidence for a possible localization of functions.

But what actually follows from Penfield's and similar experiments? It is tempting to infer localization from causal production or even to identify experiences with circumscribed brain processes, but it is, however, misleading. For even the stimulation of my foot by a needle produces a sensation of pain—nevertheless, this would not cause any brain researcher to localize the pain in the pain receptors of the skin. Pain sensation is the *integral reaction* of the living being to a peripheral stimulus, for which, undoubtedly, the activation of certain neuronal networks is also necessary.

It is in principle possible that that the same pain could be produced by the direct stimulation of the somatosensory cerebral cortex or the insula (Selimbeyoglu & Parvizi 2010). This does not, however, change the fact that the pains, in both cases, represent expressions of life, that is, *reactions of the whole organism.* The pain is experienced as suffering, it is accompanied by tensing in the body, defense movements of the foot, and an expression of pain in the face, as well as with an activation of the sympathicus system, that is, a stress reaction of the organism—all that is the pain. If it is thus not situated in the skin receptors, what speaks for localizing it in certain centers of the brain?

One possible argument is that the sensation of pain in the periphery can be suppressed by a blockade of nerve conduction, so that it is no longer felt, and can thus not represent the correlate to the sensation of pain. However, the same applies to any region of the brain. If a sufficient number of its neuronal connections are severed, its stimulation can no longer produce any sensation. Hence, a certain, sufficiently extended totality of brain activities in connection with the organism is necessary so that we can experience pains.[23] That is why those experiences cannot be localized at their triggering point and are not "identical" with certain neuronal processes. The temporal lobe does not contain any memories or sensations of smell, nor does the insula have any pain sensations, even if they can be provoked there by an electrode. Only the living being as a whole has memories and sensations.

This leads to the following conclusion, in agreement with Rockwell's (2005) account: a pain in the foot is not caused by an unconscious signal that travels up the leg and transports "information" about the event into the brain. Instead, the pain should rather be regarded as " … a network property that arises out of the relationship between the nerves in the foot, the spinal cord, and the various

[23] Selimbeyoglu and Parvizi (2010, 9) come to a similar conclusion: "Today, the phrenological notion is outdated […] perceptual and behavioral phenomena induced by electrical charge delivery to a brain region are most likely due to change of activity in a network of brain areas (including subcortical regions) rather than the excitation or inhibition of a blob of cortical grey matter per se" (Selimbeyoglu & Parvizi 2010).

neuronal ensembles in the cranium" (Rockwell 2005, 32). Whether the stimulation occurs in the foot or in the brain, in each case it means a reconfiguration of the *whole nervous system* which embodies the pain. Similarly, as we will see further later in this book, consciousness is not in the head, but spread over the whole body, and it is only modified, not "brought forth" by the local stimulation.

We see that increasing research into the functional specialization of the brain is not suited to supporting a localization of consciousness as such. The decisive reason for this is that it represents an integral activity of the organism, which, as we will still see more closely, requires continuous embedding in an environmental context. Granted, partial functions of consciousness can to a certain extent be assigned to certain specialized regions, damage to which then also results in the failure of the function. However, every theory which views consciousness as being assembled from localizable individual functions or modules incurs the problem of how these individual functions are to be integrated into a united activity—a question which is mirrored in different variations of the "binding problem." The entire project of the spatialization and materialization of consciousness all too easily loses sight of its object because of looking too closely at it, thus ending up with only fragments. Hence, what Georg Christoph Lichtenberg wrote at the end of the eighteenth century about the attempt of the anatomist von Soemmerring to localize the soul in the ventricles of the brain is still valid today:

> If I, when viewing the setting sun, take a step towards it, I come closer to it, little and all as it may be. In the case of the organ of the soul, this is quite different. Indeed, it would be possible, by means of coming exaggeratedly close, such as with the microscope, to once again distance oneself from what one can approach. (Lichtenberg 1973, 852; own translation)

How far we must step back to set eyes upon the locus of consciousness still remains to be investigated.

2.3 Third criticism: the powerless subject?

2.3.1 The unity of action

In the first step of the criticism (2.1), it was explained why subjectivity and intentionality cannot be completely reduced to physical descriptions. In a further step (2.2), we investigated the mereological and localization fallacies, to which an identification of the subject with the brain leads. A third question remains to be addressed. A reductionist neurobiologist could argue: "Sure enough, consciousness is real and possibly not completely reducible. However, it is certainly *produced by* the brain. That is also why the brain possesses reality to a greater extent than consciousness. It is the *actual* reality. And because

this reality is of a physical nature and, as a result, subject to physical principles, subjectivity itself cannot have its own effectiveness in the world." We may well believe that we ourselves direct our thoughts and actions, in reality, however, they are designed by neuronal systems, and they surface in consciousness like film scenes which a projector on our back casts on the screen.

By this means, we arrive at the discourse about free will, which has been debated for years. Indeed, it is surprising that, of all things, the human brain is called as the crown witness of determinism. For it is precisely the brain that is the organ whose growing complexity in the course of evolution has relaxed the rigid stimulus–response mechanism, thus enabling organisms to attain increasing degrees of freedom—seen from that point of view, it is the organ of freedom. We talk, for example, in psychiatry about a lack of freedom, above all in the various impairments or dysfunctions of the brain. Patients with frontal brain injuries suffer from aimlessness and a lack of initiative; they can no longer maintain a directed intentional arc, spanning longer stages. Patients with Tourette syndrome are compelled to make spasmodic movements or to express swear words, and are unable to restrain themselves. People with compulsions cannot help doing things which they themselves find meaningless, or think what they do not want to think. Schizophrenic patients even experience their actions as being directed by foreign powers. In all these cases, it is rather *disturbances* of brain functions that restrict the patients' freedom or dictate to them what they must do.

It is, however, precisely this, according to the opinion of some neurobiologists, that applies to us all: brain processes work deterministically, and we cannot do otherwise to what our brain determines. In fact, decisions are ultimately directed by unconscious emotional processes in the limbic system, and the actions are then triggered by the premotor areas of brain, before the person becomes conscious of this. Thus, the brain only deludes us into believing we are acting and responsible persons, whereas we can in fact only ratify its decisions in hindsight.

> [O]ur actions are clearly the result of a causal chain of neuronal activity in premotor and motor areas of the brain. […] although we may experience that our conscious decisions and thoughts cause our actions, these experiences are in fact based on readouts of brain activity in a network of brain areas that control voluntary action. (Haggard 2011, 404)
>
> [O]ur brains have to function as efficient, unconscious computers that nevertheless make rational decisions. (Swaab 2014, 331)

Although published over 30 years ago, Benjamin Libet's demonstration of a preceding readiness potential in the brain, in the case of subjectively experienced arbitrary movements, still functions as an *experimentum crucis* for the

neuroscience of voluntary action (Libet et al. 1983, Libet 1985). In this study, test persons were asked to wait for the impulse or "urge" to move a certain finger, and then, to state the point of time of this impulse, with the help of a rotating clock hand. EEG activity was measured at the same time, showing the emergence of the so-called readiness potential over the supplementary motor cortex 1 second or more before the actual movement, and about 500 milliseconds before the stated impulse to move. This seemed to demonstrate that action is prepared and triggered by the brain even "before you know it," at least challenging any versions of free will where intention occurs at the beginning of the decision process.

The deterministic interpretation of this experiment has frequently been criticized, above all, because it isolates human action experimentally from its intentional context and restricts it to the level of accidental movements.[24] It seems, to say the least, adventurous that the denial of free will should be based on an experiment which certainly depends on the voluntary participation of test persons, who would never have moved their finger without their consent. This preceding component, that is, the actual process of deliberation and decision is not included in this experiment at all. It thus disassembles the temporal and meaningful unity of forming one's will and acting on one's will, with the result that a final, artificial "moment of decision," a "tug of will" is created. Similarly, all further experiments on brain and volition have so far only dealt with decisions made in time frames of seconds and on extremely simple actions such as moving a finger.

> Moreover, an experiment carried out by Herrmann et al. (2008) rather suggests that the readiness potential may reflect an unspecific anticipatory stance. In this study, test persons carried out a *choice reaction* task: depending on geometrical figures presented to them at the last moment, they had to choose between pressing either one of two buttons. This was preceded by readiness potentials too, however, *before* the presentation of the respective picture, thus at a time at which the choice between the buttons could not have begun in the brain. Thus it seems likely that the readiness potential serves the general preparation of expected movements, corresponding to what Jeannerod (1997) has termed "motor imagery," but does not yet determine the final action.
>
> Libet's paradigm has meanwhile been further developed into *action prediction* by applying massive computational technology to whole-brain fMRI scans. Also using a choice task, Haynes and his group were able to predict with 60% accuracy whether subjects would press a button with their left or

[24] See Gallagher 2005, 237–240, for a critique.

right hand up to 10 seconds before they became aware of their choice (Soon et al. 2008). This seemed to question the idea of conscious decision-making. However, a more recent study by the same group confirmed Libet's initial assumption that a conscious *veto* is still possible even in the last fraction of a second: while the computer tried to predict their actions from brain activity, test subjects were able to stop their already initiated action until up to 200 milliseconds before the actual movement (Schultze-Kraft et al. 2016).

We have already pointed out the implicit dualistic preconditions of the neurobiological position (see 1.5). This also applies to the arguments against free will: they are based on the fiction of a Cartesian ego, separated from its body, its feelings, and its enactment of life, which reaches a decision in unlimited arbitrariness and then imposes its execution on the body. The effectiveness of this fictitious ego is then declared refuted by referring to the closed causal chain of bodily processes. Consciousness always comes too late compared with its neuronal construction mechanisms. The physical world leaves no scope for the causality of the subject. Consequently, decisions and actions ought to be ascribed to the brain.

Such argumentations are basically subject to the criticism regarding the mereological fallacy. Brains decide just as little as they are in the position to act. Indeed, attributing decisions to brains also negates the concept of decision itself (Fuchs 2007a): a computational, neuronal process as such, regardless of whether it proceeds in a strictly deterministic, probabilistic, or indeterministic way, is incapable of grasping alternative possibilities *as possibilities*. Indeed, it is even unable *to grasp the future*. That is why it is no more a process of decision-making than a cube falling or the function of a random generator.

The term "readiness potential" does not mean that the brain or the motor cortex could actually be "ready" or "prepared" for something to happen. This readiness can only emerge with conscious life, for only consciousness is able to integrate time into a span that includes the immediate past, present, and future. This integration has been famously described by William James (1890) as extended or "specious present," by Henri Bergson (1950) as "duration," and by Husserl (1991) as "inner time consciousness." To explain it briefly: the mere succession of conscious moments, as such, could not establish the experience of continuity. It is only when these moments mutually relate to each other in a forward and backward directed intention that the sequence of experiences is integrated into a unified process. Husserl conceived this as the synthesis of *protention* (indeterminate anticipation of what is yet to come), *presentation* (primal or momentary impression),

> and *retention* (retaining what has just been experienced as it slips away). This can be illustrated by a melody or a spoken sentence: we hear the current tone (presentation), but are at the same time still aware of the tones just heard (retention), and vaguely expect the continuation of the melody (protention). Consequently, what is perceived is not a sequence of single moments but a dynamic, self-organizing process, which integrates the tones heard into a melody, or words into a sentence.[25] From this follows that being ready or prepared for something, or anticipating the next-to-come, is only possible for a conscious living being. Indeed, to anticipate the not-yet and to retain the no-longer is one of the most fundamental functions of consciousness.

A fortiori, the anticipation of possibilities *as possibilities* is only available to a human being who finds herself in *future-oriented life conduct*, who disposes of *embodied capacities of action* and who can counterfactually also *imagine the not-being*—"to do or not to do?" is the question at every decision. Comprehending the alternatives *as alternatives* (left button or right button?) in the first place is even the precondition for all so-called decisions in the above-mentioned experiments. If, however, this subjective perspective is eliminated as illusory, then there are no alternative pathways of events; the world runs as it runs, and, consequently, brains decide nothing. Apart from this, psychology has always been aware that not only conscious and rational considerations are included in the subjective decision-making process, but rather also unconscious or partially conscious motives, dispositions, and tendencies. This does not change the fact that every decision needs anticipation and thus, consciousness.

The same applies for the concept of action. We can only speak of actions (in contrast to events) if there is a person acting, and this is the complete human being. Monica goes to school—not her Ego, her brain, or her legs. If Monica moves her legs for this, they usually do that by themselves, and there is no need for a willed decision (it suffices that she wants *to go to school*). Should Monica have the idea of moving her legs intentionally and in a targeted manner, as the Libet experiment requires it of the test persons, her legs will certainly obey her. Nevertheless, this particular instrumental relationship, which the human being can have to her body, does not produce a bodiless "Ego" or an ominous

[25] Perceptual experiments on the so-called flash-lag effect also demonstrate that we are slightly ahead of the present: if subjects are watching a continuously moving object, and a sudden flash is presented at the exact location of the object on its trajectory, the subjects erroneously see the object as having already moved past this point (Changizi et al. 2008, Nijhawan 2008).

"will" which gets the body moving from outside. Monica would not know at all how she should do that—"to move her leg," like she would move a plate out of the cupboard. She remains, also with intentional movements, an embodied being which *moves itself*—and does not transport its own legs, like two pieces of wood, from here to there.[26]

Now if specific motor readiness potentials emerge in Monica's brain shortly before she sets off, she, of course, does not become an automatic machine or a marionette of her brain. Monica could, for example, have come to the conclusion rather to play truant and to go swimming. As soon as she turns this decision into action, however, *precisely the same* readiness potentials would appear in her motor cortex. These brain activities are therefore necessary, and at a very late stage also sufficient, conditions for Monica's *muscular movements*, but are not sufficient for her *future-directed action*. For the action of going to school is, undoubtedly, a completely different action than playing truant, although they both use the same muscles and motion sequences. What the neurobiological description explains is therefore, at best, a body movement in the sense of a physiological event. In other words, it explains the *proximate* or subordinate causes of the action. To explain the movement *as action*, however, a knowledge of Monica's motives, thoughts, wishes, and aims is required—that is, thus, a quite different, namely, psychological, teleological, or intentional description. Physiological causes are completely irrelevant for the question of the *meaning* of an action. Of course, too, these subjective phenomena do not exist in a transcendental world of the mind; they are, rather, just like Monica's ability to go, manifestations of her embodied subjectivity. Hence, if one wishes to give the cause for the action *as* action, it can, therefore, neither lie in an Ego or will, nor in the brain, but rather in the complete human being with all his or her mental and bodily capacities.

2.3.2 **The role of consciousness**

Of course, one can further radicalize reductionism and can award subjectivity a merely epiphenomenal status also in the processes of consideration, evaluating, and deciding. The question is therefore whether the process of the subjective assessment of possibilities *co-determines the result*, or whether it is only a powerless mirroring of physical processes. If subjective experience in fact remained without consequences for the course of the world itself, this would indeed strike at the heart of the idea of personal freedom and agency. Is it then crucial that I seriously consult with myself about what I should do in a certain situation? Does it make a difference in the world? Would we really be able to act otherwise?

[26] This corresponds evidently to the conception of Aristotle who spoke of living beings as "self-moved."

If it is true that we do not find possibilities, evaluations, reasons, and, finally, decisions in the physical world, then it does make a difference in fact. For it means that the processes of deliberating, evaluating, preferring, and deciding cannot completely be reduced to physical-chemical laws. That brain processes are not solely determined by such laws can easily be seen, as the brain is essentially shaped by cultural, ideational, and symbolically mediated influences. For example, what counts as a logically valid inference or what the result of "$x=\sqrt{16}$" is, is not determined by natural laws of physics. So if we find "$x=\pm4$" as the solution of the equation, its correctness does not result from physical or neurophysiological but from mathematical laws. The brain is only a highly malleable carrier medium, which is capable to adopt such general laws. Such shaping of neural dispositions, however, is crucially mediated by subjective experience; we will come back to this in Chapter 6.

Now, the shaping of the brain by means of language, ideas, and culture is commonly also conceded by neuroscientists. This, however, is assumed not to change anything about our being completely physically determined: in that case, it is argued, functional equivalents of meanings and cultural programs become part of the neural algorithms, for instance, equivalents of mathematical, logical, or moral rules. But it is still the brain that carries out these programs, calculates, thinks, and "decides," since it was programmed in this and not another manner. Subjectivity and conscious experience, however, are assumed not to have an influence on the process of deliberation:

> The *sense* of will is an invention of the brain. Like so much of what the brain does, the feeling of choice is a mental model—a plausible account of how we act, which tells us no more about how decisions are really taken in the brain than our perception of the world tells us about the computations involved in deriving it. (Blakemore 1988, 272)

A central argument against such a position is based on the theory of evolution: why should subjectivity and consciousness have evolved at all? What is the point of investing such developmental efforts and energy into a phenomenon without any significance and consequence, a systematic self-deception of billions of living creatures?[27] If the brain functions perfectly well without an ancillary support of consciousness, then there seems to be no causal role for conscious processes that could improve the odds of a living being's survival.

In his account of consciousness, neuroscientist Edelman explicitly poses the question whether phenomenal consciousness has causal efficacy and thus an

[27] This kind of objection against epiphenomenalism was already put forward by Puccetti (1974) and Popper and Eccles (1977).

adaptive function (Edelman 2004, 76–88). Granted, he argues, with certain neural processes, the simultaneous property of consciousness is given in a no further deducible manner—a "phenomenal transform" of the "dynamic core" (see 2.2.2), including "what-it-is-likeness" and qualia. However, the causal closure of the physical world demands that it is not the phenomenal experiences C, but only their carrier processes C′ which can cause physical effects. These processes were selected for by evolution in order to enable efficient planning and acting, and it is they that realize causal links. The phenomenological transform only serves as a "reliable indicator of the underlying causal C′ events" for the individual (2004, 79).

Now Edelman himself does not seem entirely sure what purpose this indicator might serve if the conscious individual is nothing but a powerless accompaniment of their neurons and, for this reason, he adds another function: consciousness, at least, enables higher animals to communicate to others the states of their C′ brain regions:

> Animals so evolved would communicate efficacious C′ states in terms of C. C, after all, is the only information available that reflects C′ states to each animal and to others. (2004, 81)

Of course, Edelman has to concede that the dynamic core as a carrier of consciousness will already have developed in species "without extensive communicative abilities" (p. 81). Therefore, the only option left is to conceive of C as an "epiphenomenon" (p. 85) that is necessarily linked with C′ processes, without itself having a function. Nonetheless, Edelman finally states that "the phenomenal transform is an elegant means of conveying the integrated states of C′ on a first-person basis" (p. 86). But which function does this elegance fulfill? The claim remains tautological, for "conveying C′ states on a first-person basis," in the final analysis, means nothing else than transforming them into phenomenal experience. So in that case, phenomenal experience is good for phenomenal experience.

Here we encounter once again the basic dilemma of neurobiological approaches: the more complete the alleged physiological description of the neural foundation of consciousness, the more precarious the question of the function of consciousness itself becomes. As Hans Jonas has pointed out, it becomes "a dead-end alley off of the highway of causality, past which the traffic of cause and effect rolls as if it were not there at all" (Jonas 1966/2001, 128). More so, it becomes one of the properties that natural science wanted to eradicate from its world, namely a "*qualitas occulta*," a hidden, unprovable property of certain material processes that is manifested in no effect. Hence, there is no way around the insight that if we do not want to buy into the ontological as well as the

biological absurdity of an inconsequential subjectivity, we have to conceive of the brain in a manner that it cannot only be shaped by social and cultural influences, but also be *currently* integrated into the superordinate conscious enactment of a human being's life.

We have already been able to ascertain in various ways which fundamentally novel phenomena appear in the world with the emergence of consciousness. I summarize its most important dimensions as follows:

- The integration of the living being's sensorimotor interactions with the environment into an *intermodal action space* ("*sensus communis*"), allowing for skilled coping with environmental affordances and opening up possibilities for action.
- The *intentional and affective directedness* of a living being towards relevances and meaningful situations in its environment; that is to say, consciousness is teleological, oriented towards goals and purposes.
- The *integration of experience over time*, in the sense of being directed towards the immediate future and its possibilities (protention) as well as retaining past experiences (retention)—in other words, the temporal coherence of consciousness.
- The awareness of *alternatives of action* offering themselves in a given situation, in human beings also including counterfactual imagination ("as if").
- Last not least, the self-experience of the living being in relation to the environment, that is, a basic sense of *self-awareness and self-affection*, integrating the organism's current overall state with regard to its own self-preservation. This integration also manifests itself in the spatially extended and yet indivisible unity of the subject-body (see 1.2.2).

All these phenomena and properties are nowhere to be found in the physical world: neither a unified action space filled with qualitative affordances, nor an intentional and affective directedness, nor an integration of time, nor finally the dimension of self-awareness, which turns higher animals into centers of their own world. *Unlike physical mechanisms, consciousness is not analyzable into distinct spatiotemporal components; it covers space, time, and the body.*

> To demonstrate this with regard to temporal integration: physical processes, including neural processes in the brain, are always only present, irrespective of how complex they may be. They are never more than *linear sequences of events*, at any time restricted to the current moment, without any anticipation of a future (physiological control loops and even "feedforward" mechanisms cannot actually "anticipate" anything), or a memory of the past. It is only the overarching temporal continuity of consciousness (see 2.2.1) that allows higher animals to grasp the possible future, in particular to anticipate possible action.

> Although recent neurocognitive theories posit the brain as a "predictive organ" or "prediction machine" (e.g., Downing 2009, Clark 2013, Hohwy 2013), this should not blind us to the fact that brains are neither in the condition to advance hypotheses about possible events nor to make inferences about remote objects or predictions about the future—simply because they are not "ahead of themselves" and therefore *unable to anticipate what is yet-to-come, even less to grasp the future as such*. There may well occur an alignment of predisposed excitation patterns and incoming stimuli in the dynamical state space of the brain, in the sense that "forward models" are either matched by the input or not. But this is not principally different from correction mechanisms in "target-seeking" missiles; it means neither a "confirmation" nor a "disconfirmation" of hypotheses or anticipations. No matter how important stochastic (Bayesian) adjustment processes may be in the brain's processing of incoming stimuli, a "predictive brain" as such does not exist.

Given the irreducible integrative properties of consciousness, it seems nearly absurd to assume that this multidimensional integration, and with it, the appearance of a fundamentally novel phenomenon in the world, should have remained without consequence for the behavior and the adaptation of living beings which dispose of such a function. On the contrary, over the course of evolution, the brain has developed as an organ whose complexity enabled the emergence of feeling, emotion, thought, and volition, and which became the crucial (though not sufficient) basis of integrative conscious experience. In this way, the developing brain allowed for ever greater degrees of freedom of living beings and multiplied their scope of choice and action—up to the possibility of free deliberation and decision in human beings.

Thus, the brain is rather an organ of freedom than of necessity. There are neural processes that can function, so to speak, as a "matrix" for motives, considerations, imaginations, and evaluations, no less than for mathematical or logical laws. Neural conditions of consciousness do not exclude freedom, but are its conditions of possibility—though it is only consciousness itself which is able of envisaging possibilities as such. Hence, the alleged causal closure of the physical world should not blind us to the particular possibilities of emergence and "downward causation" that made their appearance with living beings, and which may also enable a consistent account of embodied human freedom. We will return to this issue in Chapters 3 and 6.

2.4 Summary: the primacy of the lifeworld

In this chapter, the idea that subjectivity could be reduced to the description of neuronal processes was criticized and refuted. The characteristics of phenomenal

consciousness, especially the subjectivity of experiential facts, the phenomenon of intentionality, and the integration of time, cannot be sufficiently explained by the description of correlated physiological events. Moreover, the attempts at reduction run into category mistakes which were analyzed as the mereological and localization fallacies. Finally, the claim that processes of consciousness only possess an illusory efficacy leads to the aporia that their appearance and function in evolution become a riddle. In contrast, it was shown that consciousness enables an integration of space, time, and self that is not found in the physical world and multiplies the possibilities of living beings to cope with the environment and to preserve themselves.

Following on from the "Introduction," I would now like, at the end of this first part of the book, to grasp the problems posed by neurosciences at their root and will additionally use a culturalist approach, as it was developed by Janich (1996) and Hartmann (1996, 1998). My thesis reads as follows: the problems of the relationship between brain and mind, as they present themselves today, emerge from a *short circuit* between the level of natural scientific, in this case, especially neurobiological constructs, and the level of intersubjective, lifeworld experience, from which the neurobiological special practice has developed and with which it remains always bound.

The basic paradigm which directs the cognitive neurosciences is, in the last analysis, a *metaphysical realism*: there is an objective, material world "out there" which is independent of our process of observation and of our anchoring in the lifeworld, and of which there must, in principle, be a complete, and, in fact, *physical* description (even if this description has to use certain constructs and we can only approximate completeness). If we had this complete description, it would include everything that happens in the world, that is, also *our experience and observation of the world itself.* In other words, it would have to include all that could be known about consciousness and its contents. Otherwise consciousness would be an additional, non-natural property of the world, which would contradict the precondition.

The basic problem of this approach lies in its manifest, though mostly not comprehended, circularity. It is based on the assumption that there could be a position of observation and recognition beyond our lifeworld experience which is, however, always presupposed with the observation. Independently of this experience, physical objects cannot be identified at all. What makes up a human being, a brain, neurons, molecules, or atoms can only be gathered from our common prior understanding or from conventional agreement. Metaphysical realism or physicalism is thus incoherent insofar as it overlooks its own dependence on the intersubjectively constituted lifeworld. This lifeworld is based on the basic relationship structure "We–It"; that is, as members

of a community of interaction and communication, we are jointly directed to objects in our environment. The *perspective of the participant*, that is, the "we"-perspective of the first person plural is the primary and permanent basis for the scientific observational or third-person perspective. It follows from this that a nature regarded purely physically, in which no subjects occur, must always remain a theoretical construct, from which consciousness and intersubjectivity cannot be deduced.[28]

Neurobiology is primarily a highly specialized form of common practice arising from the lifeworld. "The lifeworld includes everything we can speak about in pre-scientific terms: fellow humans, cats, sunflowers, stones, weapons, cathedrals, but also sounds, afterimages, thoughts, memories, hunger, happiness and fear" (Hartmann 1998, 322; own translation). However, initially it does not contain any constructs such as atoms, molecules, or action potentials. Within the lifeworld, human beings form cultural, linguistic, and action communities, among them also special practice forms such as the natural sciences, which raise the perspective of the observer to its methodological ruling principle. In that way, they cut out certain quantifiable and objectifiable areas from the phenomenal lifeworld, in the way described in the "Introduction." In order to describe the structures of the section of reality they choose, they develop certain terminologies, and, in due course, certain constructs (atoms, electrons, waves, potentials, fields, etc.), which serve to explain the processes observed and which, in connection with certain laws, are of high prognostic, and thus also practical value for the community. In this way, methodical norms, such as the causal principle, which were initially only research directives, gain increasing undisputed, indeed metaphysical status (such as "universal determinism").

The "second naturalistic fallacy"[29] consists, according to Hartmann, in the fact that the structures and processes postulated on the construct level are now

[28] This is in line with Merleau-Ponty's argument: "For what precisely is meant by saying that the world existed before any human consciousness? An example of what is meant is that the earth originally issued from a primitive nebula from which the combination of conditions necessary to life was absent. But every one of these words, like every equation in physics, presupposes *our* pre-scientific experience of the world, and this reference to the world in which we *live* goes to make up the proposition's valid meaning. Nothing will ever bring home to my comprehension what a nebula that no one sees could possibly be. Laplace's nebula [or today, the big bang, T. F.] is not behind us, at our remote beginnings, but in front of us in the cultural world" (Merleau-Ponty 1962, 385).

[29] The "second," because the term "naturalistic fallacy" is already used to describe the deduction of an "ought" from an "is," that means, drawing ethical conclusions from natural facts.

increasingly pushed *underneath the lifeworld experience* and, in the long run, hypostasized as actual reality:

> A knife consists of a blade and a handle, the material of the blade is an alloy which consists of molecules which are a combination of atoms, which, in turn, consist of even more minute particles—all just a matter of looking "ever more closely." It is overlooked here that the construct objects, in contrast to the objects on the phenomenal level, are not accessible independent of the theories in which they arise. (Hartmann 1998, 326)

This gradual substitution of the phenomena by quantifiable constructs remains unproblematic for the primary, that is, inorganic and mechanical objects of the natural sciences. It already becomes, however, reductionist for the phenomena of life as these presuppose complex or holistically structured and, thus, *macroscopic* bodies; they disappear from sight in the course of ever progressing division. This approach must all the more remain reductionist in the face of the phenomena of experience and consciousness because these per se evade the objectifying perspective. According to the fallacy of the ontological hypostasizing of the constructs, physical description shall now apply universally, that is, capture all conceivable aspects of reality. The lifeworld must thus be reconstructed from the constructs: a dog barking happily then consists of certain collections of organic molecules, and his barking can be explained from genetic programs. The performance of Mozart's "Requiem" consists of transitory fluctuations in air pressure in the surroundings of human beings and the heard melody is explained from the firing of neurons in the brain of the listener.[30]

This naturalistic fallacy is also the basis of all mereological and localization fallacies in the neurosciences. Their belief in an ultimately valid material reality and its lawfulness, existing independent from any observer, is drawn from physicalism. According to it, the subjective worlds must be grasped as constructs which are produced by the physics of the brain. The general, naturalistic short circuit between the level of physical-chemical substructures and the level of the lifeworld then becomes the short circuit between brain and mind, or brain and subject.

Of course, quantum physics has long since shown that it is no longer possible to exclude the point of view of the observer, particularly in exploring the

[30] "The physicist's atoms will always appear more real than the historical and qualitative face of the world, the physico-chemical processes more real than the organic forms [. . .] as long as the attempt is made to construct the shape of the world (life, perception, mind) instead of recognizing, as the next source and as the ultimate court of appeal in our knowledge of these things, our *experience* of them" (Merleau-Ponty 1962, 20; translation slightly modified according to the French original, T. F.).

elementary processes, whereby the allegedly solid ground for reductionism becomes shaky. The physicist is left with neither fixed "building blocks," nor completely objectifiable "facts," from which the world could be assembled as from a construction kit. The idea of matter in the sense of interacting pieces such as billiard balls is long since outdated. The processes of the material world are no more directly given than other aspects of reality. Since, consequently, the neurosciences are also dependent on the observer, they cannot explain observation itself as a product of their object.

The basic thesis of physicalism that all areas of reality can be described either by physical concepts and laws or that their own local theories can be reduced to physical theories is untenable as well. The practice of empirical sciences, such as biology, psychology, or sociology, more than underlines that their explanations of the phenomena in their particular branch have nothing at all to do with physical theories. The prerequisite that their explanations *do not contradict* basic physical principles (thus, e.g., no non-physical natural powers are introduced) suffices for them. However, the description and explanation of phenomena in accordance with physical laws does not mean that the explanation itself can be a physical one. The happy barking of the dog cannot be satisfactorily elucidated either by the biochemical analysis of motor endplate activation in his vocal muscles or by a physical description of the atomic or subatomic processes in his brain. Physical or physiological descriptions cannot explain the Russian Revolution, just because the people and things involved in it consisted of matter and cells. Admittedly, the Communist Program did not exist without material carrier substances, for example, in the form of black lettering on newspaper pages, or in the form of certain excitation patterns in Lenin's brain. Nevertheless, it can at best be neurobiologically sufficiently explained why Lenin was no longer able to pursue his program in his last years of life—namely because of several strokes he suffered.

The basic naturalistic fallacy on which the search for the substrata of consciousness in the neurosciences is based has, as of now, not been worked out. Even if the concept of "social cognitive neuroscience" (Cacioppo et al. 2002, Decety & Ickes 2011, Cozolino 2014, and many others) is meanwhile firmly established—the neuro- and cognitive sciences can only become *social* neurosciences when they incorporate not only the observer perspective, but also the participant perspective in their concepts and research. The latter is, in contrast to the observer perspective, the actual social perspective in which people recognize one another as persons and, as such, communicate with one another. Their experiencing, perceiving, feeling, and acting can only be captured from this perspective and then, with certain restrictions, also be correlated with neuroscientific findings. If someone does not know what "seeing" is, and if they cannot

communicate with other seeing persons, they cannot perform any neurophysiology of visual perception. The very constitution of his objects demands that the neuroscientist takes the perspective of the participant. Moreover, scientific discourse, too, presupposes that the persons involved recognize one another as judicious and capable of freely reaching agreement. Hereby, they do not relate to a construct level of physical descriptions, rather they relate to a common lifeworld as their meaningful context and horizon, which is represented by cultural patterns of interpretation, handed down by tradition. "Without intersubjectivity of understanding, there can be no objectivity of knowledge" (Habermas 2004, 885).

Thus the lifeworld experience gains a weight which puts the complete burden of proof on its denial. The special practice of brain research is justified as long as it does not lead to hyperbolic conclusions, intended to highlight lifeworld experience in its entirety as secondary or even illusory. Whoever would wish to undermine this experience by physiological constructs or brain-generated self-models, cannot invoke scientistic doctrines such as the complete physical reducibility and causal determination of all phenomena. In fact, it is rather the other way around: the models of brain research, as soon as they transgress the level of pure anatomic and physiological research and touch the field of subjectivity and consciousness, must orient themselves primarily to plausibility for our experience—thus, for example, stating which neuronal conditions exist for this experience—and not to a physicalist world view, in which colors, tones, feelings, actions, and, above all, subjects do not occur a priori anymore.

It follows that a theoretical model which is suitable for an adequate interpretation of the neurobiological data and insights must start from the perspective of the first and second person, that is, from the self-experience of living persons and must return to it, without losing it on the way. On this assumption, in what follows I shall develop a view of the brain compatible with lifeworld experience.