2 Is free will an observer-based concept rather than a brain-based one? A critical neuroepistemological account

Georg Northoff

ABSTRACT

How are the brain and our observation of it in neuroscientific investigation related to each other? I here distinguish between brain-based and observer-based concepts. Brain-based concepts mean that they are very much in accordance with the way the brain functions and processes neural activity independent of our observation of it. In contrast, observer-based concepts refer to the dependence of our observations on the observer and his or her particular experimental (and technological and other) requirements. Since in observer-based concepts the observer intrudes into the observations and ultimately into the brain itself, I also speak of 'observer-related intrusions'. I distinguish between extrinsic and intrinsic observer-related intrusions: extrinsic observer-related intrusions can in principle be avoided and minimized while intrinsic ones cannot in principle be overcome. This is so because they concern, as I assume, intrinsic design features of the brain, and do therefore define the brain as brain. I conclude that the intrinsic observer-related intrusions pose 'neuro-experimental and neuroepistemological constraints' on our possible knowledge of the brain. Extrinsic observer-related intrusions pose methodological fallacies that are to be avoided, as, for instance, the projection of the concept of free will onto the brain.

Introduction

There is an intense debate about free will in current philosophy and neuroscience. Philosophical discussion often focuses on the metaphysical question of whether human agents can choose and act freely in light of causal determinism. Neuroscientific discussion focuses on the neural mechanisms underlying the will and what they imply about whether or to what extent it is free. If these mechanisms are part of a process in which they do all or most of the causal work in producing actions, then this seems to rule out free will. On this view, our behavior is controlled

not by what occurs at the mental level of beliefs, desires and intentions but instead by what occurs at the neural level (Farah 2005; cf. Walter 2001). The main question is whether the demonstration of neural activity is a sufficient condition to opt for neural determinism and the implication that we lack free will.

My approach in this chapter will be different. Rather than taking the concept of free will for granted and then projecting a definition of it onto the brain, I opt against such a 'mapping strategy'. Instead, I focus on the epistemological presuppositions on which the application of the concept of free will to the brain is based. For that I will discuss at length epistemological considerations like the distinction between observer- and brain-based concepts as well as intrinsic vs. extrinsic observer-related intrusions. These imply that the current philosophical models and concepts of free will, including their mapping onto the brain, may be conceptual artifacts produced by the observer that as such are projected upon the brain. Hence, I designate free will as primarily an observer-based concept, as distinguished from empirically plausible concepts that then would need to be described as brain-based concepts.

This opens new lines of inference. Traditionally, it is assumed that neuroscientific results have direct implications for the question of determinism or indeterminism and hence whether there is free will or not. Independently of whether one opts for determinism or indeterminism, one draws a direct inference from the neuroscientific data to the conceptual question of determinism versus indeterminism. One thus implicitly presupposes the concepts of determinism or indeterminism to be brain-based, since otherwise one could not draw direct inferences from brain to concept. This changes once one associates the concepts of indeterminism and determinism and especially the apparently mutually exclusive character of these concepts being observer-rather than brain-based. Then a direct inference from brain to concept seems no longer plausible. Instead, in the first instance, one would then investigate the relationship between observer and brain rather than the one between brain and concept. If it turns out that the concepts of indeterminism and determinism can be located towards the observer end of the continuum between brain- and observer-based concepts, then any inference from brain to concept may be even more distant and problematic. If, in contrast, the concepts of indeterminism and determinism can be located towards the brain end (being brain-based to a higher degree) of the continuum, an inference from and direct relationship between brain and concepts seems to be more plausible. In other words, I claim that the possibility of drawing inferences from our current neuroscientific data to either determinism or indeterminism very much depends on the characterization of

these concepts as either observer- or brain based. The empirical plausibility of our neurophilosophical assumptions about free will may thus need to take into account the epistemological characterization of the concept itself as either brain- or observer-based.

Neuroepistemological remark: distinction between brain- and observer-based concepts

Neuroscience acquires data and ultimately facts to describe the brain (the distinction between data and facts may by itself be worth discussing from a philosophical point of view, though it is not pursued here; see Northoff 2012). Thereby it uses concepts to describe those data and facts. Usually, one would expect the concepts to correspond to and thus match the data and facts. In this case, there is a one-to-one relation with one particular datum being the only and exclusive content the concepts describe.

Life is not that easy, though, especially the life of a neuroscientist. Concepts are usually more general and vague than particular data and facts. This means that concepts usually include more than one particular content and thus by definition are general. That in turn makes them less specific when compared to data and facts. The concepts the neuroscientist (and any scientist) uses (or must use) remain consecutively unable to completely match and correspond to the data and facts in a one-to-one way. Instead, the concepts may also refer to contents other than the ones from particular data and facts in question. This implies a one-to-many relationship where one concepts stands for (or codes) many data/facts. Hence, rather than coding data and facts in either a local (even sparse) way, concepts seem to encode data and facts in a rather dense way. This means that there is almost certainly a certain degree of mismatch between concepts and data/facts.

This all sounds very philosophical, the neuroscientist may want to say. Let the philosophers discuss this, but leave me alone in generating data and facts. As noted, life is not that easy. Because of the almost certain mismatch between concepts and data/facts with the former encoding the latter in a dense way, we are prone to confusion. More specifically, we can never be completely sure (or know as the epistemologist may want to say) whether the concept we use to describe our data and facts really matches and corresponds to the latter. And where there is uncertainty, there is the possibility of mismatch and confusion. This applies to both the neurophilosopher and the neuroscientist. Given the need to start our experimental investigation with concepts and terminate it with the interpretation of the data via concepts, there is no way of avoiding concepts even in neuroscientific investigation. However, the seemingly unavoidable (or necessary) presence of concepts in the interpretation that are

often not distinguished from the one used to describe the data themselves suggests intrusion of the observer into the neuroscientific investigation. There seems to be almost no way of avoiding the observer's intrusion even in the seemingly most objective investigation. There is thus a neuroepistemological constraint, the intrusion of the observer via the use and definition of concepts, inherent in our neuroscientific and neurophilosophical methodology. The only way to minimize that intrusion may be to introduce methodological and epistemological distinctions like the one introduced here: brain-versus observer-based concepts and intrinsic versus extrinsic observer-related intrusions (see below).

How can we now describe such possible confusion between concepts and data/facts in further detail? Concepts are generated by the observer. The very same observer who conducts the experiments also needs to generate concepts to describe his or her data/facts and to formulate his or her hypotheses. Yielding hypotheses and data/facts is possible only when considering certain requirements that need to be fulfilled within the experimental context. One such experimental requirement is the careful distinction between different experimental variables that need to be treated in a segregated and independent way. This makes necessary the introduction of concepts describing these segregated and independent variables.

So far, so good. The problem starts once the very same concepts that describe these segregated and independent experimental variables are also assumed to describe the brain itself. More specifically, based on the experimental data/facts, the concepts describing the respective experimental variables are assumed to describe one-to-one the processes and mechanisms in the brain itself. Thereby the concept is supposed to match and correspond to the brain's neuronal processes as they are independent of the observer's observation. The concept is thus supposed to refer to the brain as it is by itself. However, one could also imagine instances where the concept does not match or correspond to the brain's neuronal processes and mechanisms as they are by themselves. In that case, the concept is more related to the observer and his or her experimental requirements than to the brain itself and its neuronal processes and mechanisms considered independently of the observer. This means the concept is more observer-based than brain-based. I hence distinguish between what I describe as observer- and brain-based concepts.

The distinction between observer- and brain-based concepts is not an All-or-Nothing distinction but rather a More-or-Less distinction. This means that a particular concept may be based on both the observer's experimental requirements and the brain's neuronal processes thus reflecting a continuum between brain- and observed-based concepts

(see Figure 2.1a). It may thus be just a matter of degree and balance between the two ingredients, observer and brain, that determine the concept in question. A concept is thus either more or less strongly based on either the observer and his or her experimental requirement or the brain's neuronal processes. There is thus a continuum with its two extreme poles describing purely observer- and brain-based concepts. The neuroscientist seeks of course concepts where the balance is tilted strongly towards the brain-based pole of the continuum and away from the observer-based pole.

Neuroepistemological remark: examples of observer-versus brain-based concepts

In neuroscience, we encounter several examples of suspicious concepts where the balance seemed to be more strongly tilted towards the observer than the brain itself. In the following I want to briefly mention some of them.

Central neurotransmitters in the brain are glutamate and GABA (gamma-aminobutyric acid). Experimentally, we need to segregate glutamate and GABA and correspondingly neural excitation and inhibition from each other. In order to measure, for instance, glutamate and neural excitation, we need to experimentally parse both variables from any traces of GABA and neural inhibition. Otherwise we cannot be sure whether our data really tell us about glutamate and neural excitation themselves. This means ultimately that GABA and glutamate and hence neural inhibition and excitation are treated as segregated and independent experimental variables. The designation of GABA and glutamate as segregated and independent variables occurs on purely experimental grounds and is therefore strongly observer-based. Based on the data whose acquisition presupposes such experimental segregation and independence, one would assume GABA and glutamate to also act segregated and independently in the brain itself. One consecutively assumes that certain levels of GABA and neural inhibition are necessary for a specific neuronal process. While these levels may be open to (secondary) modulation by glutamate and neural excitation, they are considered (primarily) as independent and segregated (in a constitutive rather than merely modulatory sense.

What does this imply for our distinction between brain- and observer-based concepts? This means that now the observer's concepts are transferred to the brain itself. The initially observer-based characterization of GABA and glutamate as independent and segregated experimental variables is now projected onto the brain itself and assumed to accurately

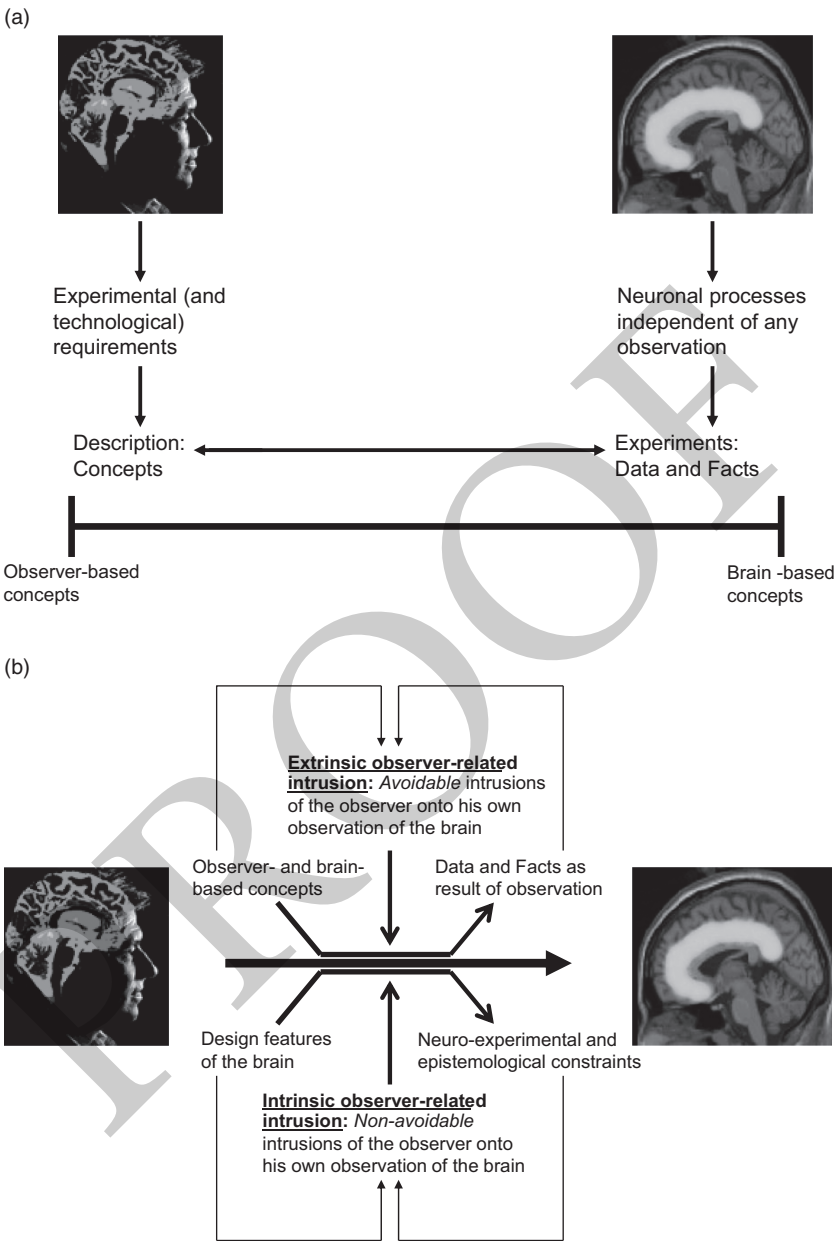

Figure 2.1 The figure describes two key features in the relation between brain and observer, the continuum between brain- and observer-based

describe its neuronal processes. In short, it is no longer regarded as observer-based but rather as brain-based.

Does such experimentally based segregation and independence between GABA/neural inhibition and glutamate/neural excitation really correspond to the empirical data? I deny that. Instead, I assume that both GABA and glutamate can be characterized by difference-based coding with each coding the relative relation between the two rather than their absolute value independent of this relation (see Northoff 2014a). This is empirically manifest in the excitation–inhibition balance (EIB) that is supposed to provide the measure for the subsequent generation of neural activity. Yet such coding of the difference to the respective other puts the assumption of (primary and constitutive) segregation and independence between glutamate and GABA into doubt. The characterization of GABA and glutamate by (primary and constitutive) segregation and independence may be relevant (and even required) for the observer (and his experimental approach). It may not apply to the brain as such independently of the observer. This means that such characterization is more strongly related to the observer and his or her experimental requirements than to the brain's neuronal processes as they are by themselves independent of the observer's observation. In other words, assuming segregation and independence between GABA and glutamate may turn out to be more observer-based than brain-based.

Another example is the distinction between different types of stimuli according to their origin in either the brain, i.e. neuronal stimuli, the

Caption for Figure 2.1 (*cont.*) concepts (a) and the distinction between intrinsic and extrinsic observer-related intrusions (b). (a) Observer-based concepts are those concepts that describe the brain but rely on us as observers and or abilities to observe the brain. Hence, we cannot be fully sure whether these concepts are related more to us as observers or the brain itself. In contrast, brain-based concepts describe the brain as it is by itself independent of our observation of it. The data reflect the brain itself rather than us as observers. I assume a continuum between brain- and observer-based concepts (lower part) with both being extreme cases in their pure forms on a continuum. (b) Extrinsic observer-related intrusions (upper part) are those intrusions of the observer into the brain's neural activity that can in principle be avoided by better experiments, etc. In contrast, intrinsic observer-related intrusions (lower part) are those that can in principle not be avoided because observation is supposed to be possible only on the basis of the brain's intrinsic features without which observation would remain impossible. Hence, observation cannot but intrude into the brain and its intrinsic design features because otherwise it would remain impossible in principle.

body, i.e. interoceptive stimuli or the world, i.e. exteroceptive stimuli. Based on these distinct origins, different anatomical structures and pathways have been assumed as it is well reflected in the radial-concentric threefold anatomical organisation (see Northoff 2014a). However, on a functional level, the distinction between the different origins of the stimuli and their respective anatomical structures seems to be blurred: this was, for instance, well apparent in the observed neural activity, i.e. functional connectivity and low- to high-frequency fluctuations, and the coding strategy, i.e. difference-rather than stimulus-based coding, that operated across and superseded the underlying anatomical structure (See Northoff 2014a). Yet this means that the distinction of stimuli according to their origin, i.e. matter of origin, may be not as relevant for the brain itself and its neuronal processes as it is for us as observers and our experimental requirements: the experimental requirement is not to confuse stimuli of different origins. Otherwise we cannot say anything about, for instance, exteroceptive stimuli and their underlying neuronal processes.

However, as important as the question of the origin of stimuli is for us as observers, it does not seem important for the brain itself. The brain seems to be more 'interested' in varying degrees of statistically based differences between different stimuli (and itself) than the stimuli themselves and their respective origins. Hence my characterization of the brain's neural operation as a 'matter of degrees and differences' rather than a 'matter of origins and stimuli'. This means, though, that the characterization of the brain's neural processing by 'matter of origin and stimuli' may be more strongly related to the observer him- or herself than the brain itself. I assume that the 'matter of degrees and differences' is tilted more towards the brain-based pole in the continuum between the extremes of purely observer- and brain-based concepts.

Let us provide a final example where brain- and observer-based concepts may be confused: the distinction between resting-state and stimulus-induced activity. Experimentally, we clearly need to segregate and delineate both since otherwise we will never be able to know the contributions of the stimulus and that of the brain itself in stimulus-induced activity. One may consecutively assume segregation between resting-state and stimulus-induced activity. Yet as the empirical data suggest, both cannot principally be distinguished from each other, let alone segregated.

Resting-state activity and stimulus-induced activity are not categorically distinct but can be distinguished from each other only on the basis of degrees. This means that the principal distinction between resting-state and stimulus-induced activity is more strongly based on the observer

than the brain itself. One may thus formulate what can be described as a 'continuity hypothesis' that assumes neuronal continuum and discontinuum between resting-state and stimulus-induced activity. If the stimulus-induced activity does not alter much the pre-existing resting-state activity, the latter may shift towards the brain end of the continuum between brain- and observer-based concepts. If, in contrast, the extrinsic stimulus and its stimulus-induced activity exert major changes to the brain's intrinsic activity, e.g. its resting-state activity, the increasing discontinuum between the two may go along with a shift towards the observer end on the continuum. Why? Because the observer is always already involved in the extrinsic stimuli he or she applies, be it directly as part of that stimulus itself, or indirectly as the applicant or cause of that stimulus. In other words, the investigation of extrinsic stimulus-induced activity may be more prone to intrusions by the observer and thus to the consecutive development of observer-rather than brain-based concepts (Northoff 2014a and 2012).

How can we escape the possible confusion between brain- and observer-based concepts? In order to shift concepts away from the observer-based pole towards the brain-based pole, we thus need to also shift our perspective. More specifically, we need to abandon our observer-based perspective and imagine how it is for the brain itself independent of our observation to generate the kind of neuronal processes we observe. We should aim to move from the observer's perspective to the brain's perspective, taken in a figurative sense because the brain itself has no perspective. Metaphorically one may consequently say that we need to replace the 'What it is like for the observer' by the 'What it is like for the brain'

Taken together, I here demonstrated various examples of concepts that seem to be more strongly based on the observer himself and his experimental requirements than on the brain's neuronal processes independent of any observation. We may want to sharpen the point even further: is an observer-free and thus truly objective (in an absolute rather than relative sense) investigation of the brain possible at all for us? Or are there always already some intrusions which we cannot avoid by all means, e.g. intrinsic observer-related intrusions? And one may argue that already raising this question is possible only on the basis of an observer. Hence no data about the brain can be interpreted without an observer and the possibility of intrusion. Diagnosis of the more strongly observer-based component in these concepts led me to search for other concepts that are presumably more brain-based. This in turn was accompanied by a shift in experimental variables and hypotheses in order to test these more brain-based concepts experimentally and lend empirical

support to them. That also makes it clear that the distinction between observer- and brain-based concepts is not a mere philosophical-conceptual exercise but highly relevant both experimentally and empirically.

Neuroepistemological remark: extrinsic observer-related intrusions

How can we be sure that the concepts we apply are more brain-than observer-based? The only way for us to know is to develop corresponding hypotheses and conduct the appropriate experiments. If the data are in accordance with the characterization implied by these concepts, the assumption of them being more brain-based may be justified. If, in contrast, the data do not support them, my concepts may turn out to be as observer-based as the ones I replaced.

If the data are in accordance with the concepts, then the latter are empirically plausible. They are thus to a higher degree based on the brain than the observer. The opposite case of no empirical support suggests that they are based more on the observer than the brain. Hence, the degree of empirical plausibility, the accordance of the concept with the empirical data, may be regarded as a measure of the degree to which the concept is more strongly brain- or observer-based. This also implies that purely brain-based concepts seem more of an ideal than reality. What further supports this view is the problem of drawing valid inferences from brain images to the data, given that there is no direct correspondence between them (Roskies 2008). To formulate it in even stronger terms, purely brain-based concepts may in principle be impossible. However, alternative experimental designs should also be applied using different experimental variables. If they yield the same or analogous results, the likelihood of both data sets being confounded by the experimental requirements of segregation and independence of experimental variables is rather low. The data may then provide an excellent basis for being associated with a concept that shows a high probability of being brain-based rather than observer-based.

How is such low empirical plausibility with a more strongly observer-based concept possible? In that case the observer and his or her experimental (and technological and other) requirements seem to intrude too much into the concept and the subsequent experimental design to yield more brain-based concepts. The observer thus intrudes into the brain and imposes him- or herself, thereby manipulating what he or she can observe from the brain's neuronal processes according to his or her own stance and needs. I call such intrusion 'observer-related intrusion'.

The concept of the 'observer-related intrusion' refers to the intrusion or imposition of the observer him- or herself into/onto his or her own observation of the brain's neuronal processes (see Figure 2.1b). In short, observer-related intrusions describe that the observer him- or herself confounds his or her own observations. Observer-related intrusions do consecutively lead to low degrees of empirical plausibility and more observer-than brain-based concepts.

How can we deal with observer-related intrusions? We can try out alternative concepts and conduct the respective experimental designs. Then we can compare the results from both experimental lines and see how much they accord with the respectively presupposed concept. The experimental line with the higher degree of accordance, i.e. empirical plausibility, may then be the one where the concept is less observer-based when compared to the respective other one. We have to try out different alternative concepts and put them to rigorous experimental testing (see chapter 3 in Northoff 2011 for such methodology, which I describe as 'concept-fact iterativity'). This means that we are not at the mercy of observer-related intrusion and that we do have methodological tools to minimize and ultimately avoid it. We can minimize the degree to which an observer intrudes and imposes himself onto our concepts. In the best case, we can avoid observer-related intrusions altogether; respective concepts are strongly brain-based approaching the one extreme end of the continuum between brain- and observer-based concepts. Since we are in principle able to minimize the degree of observer-related intrusion, I characterize them as extrinsic. The concept of 'extrinsic observer-related intrusion' suggests that the observer's intrusion and imposition can in principle be minimized and at best be avoided thus remaining extrinsic to both observation and concept.

Neuroepistemological remark: intrinsic observer-related intrusions

I assume that extrinsic observer-related intrusions can in principle be minimized and at best avoided altogether. This is possible by refining our concepts, as described, and developing better and more precise technological tools for measuring and acquiring data (as, for instance, higher resolution brain scanning). There may be instances where we remain unable to minimize observer-related intrusion. This, to be clear, does not concern the individual observer as distinct from other individual observers; rather, it pertains to all possible observers.

Let me start with Buszaki and his emphasis on rhythms and oscillations. He argues in his excellent book *Rhythms of the Brain* (2006) that

rhythms and oscillations are a hallmark feature of the brain. To prove his point, though, he would need to experimentally investigate a brain without oscillations and see whether it does not show the kinds of effects for which he assumes oscillations to be necessary. Yet that remains impossible since we cannot even imagine a brain without oscillations, let alone test it experimentally, as Buszaki himself remarks (see Buszaki 2006, 360).

Even pathological cases like schizophrenia, depression or vegetative states that may help in overcoming extrinsic observer-related intrusions do not provide an option here. Why? Because they still show rhythms and oscillations which, despite being distorted, are nonetheless present and thus not completely absent as experimentally required. There is thus a limit to the possible experimental testing that in principle cannot be overcome or avoided. More specifically, to gain experimental proof of the causal role of oscillations in specific psychological processes and even consciousness, we would need to eliminate them completely. This is not the case even in neuropsychiatric disorders. There oscillations are still present, though in a distorted way. From those abnormalities we may gain some clues about the possible role of oscillations for particular psychological processes, especially if the latter are also altered in the psychiatric patients. However, this is not sufficient to demonstrate a causal role of oscillations for the respective psychological processes but only a modulatory or correlational role. Mere modulation or correlation is not to be equated with causality since the alterations in the patients' psychological processes may be modulated by the oscillations yet caused by other neuronal processes completely different from the oscillations. Hence, neuropsychiatric disorders can help in our understanding of the brain as distinct from the observer but not overcome our epistemological limits, e.g. intrinsic observer-related intrusions, as I will call them below.

We can describe such limits in further detail. The limits consist in the fact that we remain in principle unable to prove whether our concepts of rhythms and oscillations are ultimately based on the brain itself independent of us as observers or whether they reflect how we can (and cannot) observe the brain. We are thus stuck in our own intrusion, i.e. the concept of rhythms and oscillations, without being able to ever free ourselves from them. I therefore speak of an 'intrinsic observer-related intrusion' as distinguished from an 'extrinsic observer-related intrusion'.

How is it possible that 'observer-related intrusions' are intrinsic rather than extrinsic? This amounts to the question for the different concepts associated with intrinsic and extrinsic observer-related intrusions. The concepts of rhythms and oscillations refer to a feature that characterizes the brain's designs and, even stronger, defines the brain as brain. Buszaki

cannot even imagine a brain without oscillations because otherwise he would no longer talk about a brain (at least not of a human brain) which would be meaningless. He thus considers rhythms and oscillations to be what I describe as 'design features' of the brain that as such define the brain as brain. In short, rhythms and oscillations are design features of the brain that are intrinsic to the brain and do therefore define the brain as brain.

We can describe the brain's 'intrinsic design features' in further empirical detail. The brain's intrinsic activity seems to show an elaborate temporal and spatial structure as its design feature. The temporal structure seems to consist in the fluctuations of the intrinsic activity level in different frequency ranges (from 0.001 to 60Hz). Thereby the phases, e.g. their onsets and peaks, in the different frequency ranges are somehow coupled and linked together, which seems to provide the template with a quite elaborate (not yet fully understood) temporal structure (see Northoff 2014a). The spatial structure of the intrinsic activity seems to consist in the distribution of the activity levels across different regions yielding transregional balances and thus a certain spatial pattern or structure of intrinsic activity. Moreover, it seems that temporal and spatial structures are somehow related to each other, with the exact mechanisms of such spatiotemporal coupling remaining currently unclear (see Northoff 2014a and 2014b).

How is the neural activity in this spatial and temporal structure coded? Empirical data suggest that it is not the single peak at some discrete point in time and space that is coded. Instead, what are coded in neural activity in the resting state (and also the stimulus-induced state) are the differences in neural activity between two different discrete temporal and spatial points. This means that the neural activity is not based on the single stimulus or single activity change at one discrete point in time and space but rather on spatial and temporal differences. This is what can be called difference-based coding as distinguished from stimulus-based coding (see Northoff 2014a).

I have assumed difference-based coding as distinguished from stimulus-based coding to define the brain's neural code. Since the code very much defines what and how the brain can process neural activity, it also defines the brain as brain. Since difference-based coding describes a coding strategy that results in temporal and spatial sparsening of neural activity, sparse coding must also be regarded as design feature. The same holds for predictive coding that also seems to be unavoidable once one presupposes difference-based coding.

Besides the coding strategy, other more specific design features concern the high- and low-frequency fluctuations of neural activity in both

resting-state and stimulus-induced activity. This mirrors Buszaki's assumption of rhythms and oscillations. Functional connectivity between different regions during both forms of neural activity is yet another design feature. Finally, most important, the brain's intrinsic activity, its resting-state activity, and its consecutive constitution of a spatiotemporal structure must also be regarded as a design feature of the brain without which the brain would not be a brain (at least not a human brain). 'Intrinsic feature' means here that it cannot be changed in principle by any extrinsic stimulus in the same way that the intrinsic muscle structure of the heart cannot be abolished by extrinsic blood flow Despite their differences, all these features share the idea that their absence could not even be imagined without abandoning the ground of the brain. They must therefore be assumed to define the brain as brain and are thus what I describe as the brain's design features.

We now face a serious problem. One may focus on these design features. These features may be more brain-based than other rival concepts that I regard to be more observer-based. In order to show that these concepts are brain-based, I would need to put them to experimental testing. That means that I would need to show not only that the presence of the intrinsic design features induces neuronal and phenomenal/mental effects but also that their absence makes the neuronal and phenomenal/mental effects impossible. If I were able to show the latter, I could assume that the brain's intrinsic design features are a necessary condition of possible consciousness (see Northoff 2014b).

Yet this is the point where the problems start. While I can test the effects of the presence of these design features, the experimental testing of their absence remains in principle impossible. In the same way that Buszaki cannot even imagine a brain without oscillations, let alone experimentally test its effects, we cannot imagine at all a brain without difference-based and sparse coding, a brain without functional connectivity, a brain without intrinsic activity, and a brain without spatiotemporal structure. Why? Because these are design features of the brain that are intrinsic to the brain as brain.

Let me be clear. There are principal constraints (and ultimately limits) to how far we can go experimentally. Since these principal constraints (and ultimately limits) can be traced back to the brain itself and its particular design features, I here speak of neuro-experimental constraints. These neuro-experimental constraints do constrain and limit the knowledge we can possibly acquire about the brain These epistemological constraints, i.e. the limits in our possible knowledge of the brain and its empirical function, may therefore be described as neuroepistemological constraints. Since these concern the brain's design features

and their neuronal and phenomenal/mental effects, any hypothesis is very much prone to both neuro-experimental and neuroepistemological constraints.

Does that invalidate future hypotheses? Yes and No. Yes, because ultimately one remains principally unable to know whether one is right or wrong, thus being unable to overcome intrinsic observer-related intrusion. One will therefore remain principally unable to know whether the brain operates on the basis of difference-based coding independent of my concept and hypothesis of it. No, because one can at least work on minimizing the extrinsic observer-related intrusions and develop novel experimental designs in order to put my hypotheses on more secure empirical ground. Hence, empirically one may move forward while epistemologically one remains stuck in the border territory between what can and cannot possibly be known in principle about the brain. That isn't so bad after all. Or is it?

Conclusion: free will as observer-based concept and extrinsic observer-related intrusion

How does all that apply to the concept of free will? There is intense discussion in the current philosophical debate as to whether free will is compatible or incompatible with the supposed determinism of the neuronal processes in the brain. There are different versions of compatibilism and incompatibilism in response of the idea of neural determinism and various intermediate positions between them. We did not discuss any of that here. Rather, my aim was to show that the concept of free will by itself may be more related to the observer than the brain. We as outside observers of the brain experience or assume to experience free will. However, when investigating the brain we assume determinism of neuronal states as physical states which seems to be incompatible with the assumption of free will. Is there now free will in the brain or not? If so, one would assume free will to be a brain-based concept. If not, then it is an observer-based concept. We would thus make our decision based on the kind of concept free will is dependent on our results.

We may start earlier, though. The question of free will seems to presuppose the sharp and mutually exclusive distinction between free will and no free will, between indeterminism and determinism. How does that presupposition stand to the brain and its neural operations? The brain may be neither a completely deterministic nor a completely indeterministic system. Instead, it may operate on a continuum ranging between the extremes of determinism and indeterminism. But given the neuro-experimental and neuroepistemological constraints that I have

described, the concept of free will may turn out to be more of an observer-than a brain-based concept.

There is yet another reason why the concept of free will may be more of an observer-based rather than a brain-based concept. Free will can be experienced only by persons. The brain itself cannot experience free will. Attribution of free will to the brain and specific neural mechanisms may thus lead to confusion between persons and brains implying confusion between observer and brain. The concept of free will, as person-based, is thus an observer-based concept. Characterizing free will as a brain-based concept would neglect the fact that the observer is a person, an agent who interacts with and whose actions can have an impact on the natural and social environment. Hence, it is important to consider the agent and how he experiences subjectively from the inside, in his First-Person Perspective, as well as what we as observers describe as free will from the outside in our Third-Person Perspective. We may thus need both first- and third-person approaches to properly account for free will (see also Spence 2009).

Why is all that important? By investigating such conceptual and epistemological issues, we may clear the way for subsequent experimental research into concepts that may not be suitable for inclusion in empirical investigation. Any observer-based definition of concepts that does not turn out to be brain-based has a high likelihood of not providing insight into how the brain functions and operates independently of us and our observations. In other words, by using such concepts as free will we may intrude upon the brain, yielding results that are external rather than intrinsic to the brain itself. We will produce what I described as extrinsic observer-related intrusions.

This also has important implications for claims from Libet (1985), Fried (2011), Haggard (2005 and 2008), Haynes et al. (2007) and Haynes (2010) and other neuroscientists that mechanistic processes in the brain precede our conscious intention to act and therefore rule out free will. Neuro-experimental and neuroepistemological constraints imply a gap between what we observe about the brain and what actually occurs in the brain. Because of these constraints and the difficulty of eliminating intrinsic observer-based intrusions into the brain, we can only approximate and not achieve knowledge of neural mechanisms. The argument that we lack free will and that our belief in it is an illusion appeals to recorded data about event-related potentials in motor, pre-motor and supplementary motor areas in the brain. However, if we do not have direct but only indirect access to these neural mechanisms because the data are sensitive to intrinsic observer-related intrusions, and if claims about the significance of these data are more

observer-based than brain-based, then we may have good reason to be sceptical of this argument.

There is no purely brain-based concept of free will. Claims that we have or lack free will may be more reflective of observers' projections onto the organ that is being observed than of intrinsic features of the organ itself. The continuum model offers the most plausible way of discussing the neuroscience of free will. Judgments about actions being free or unfree are not absolute but fall somewhere between observer-based and brain-based concepts along the continuum. This model suggests that appeal to neural activity alone will not establish whether we have or lack free will simply because direct inference from the brain to concepts like free will is epistemologically implausible given our apparent impossibility of excluding both observer-based concepts and intrinsic observer-related intrusions.

REFERENCES

Buszaki, G. 2006. *Rhythms of the Brain*. New York: Oxford University Press.

Farah, M. 2005. Neuroethics: the practical and the philosophical. *TRENDS in Cognitive Sciences* 9: 34–40.

Fried, I., Mukamel, R. and Kreiman, G. 2011. Internally generated preactivation of single neurons on human medial frontal cortex predicts volition. *Neuron* 69: 548–562.

Haggard, P. 2005. Conscious intention and motor cognition. *Trends in Cognitive Sciences* 9: 290–295.

2008. Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience* 9: 934–946.

Haynes, J.-D. 2010. Beyond Libet: long-term prediction of free choices from neuroimaging signals. In W. Sinnott-Armstrong and L. Nadel (eds.), *Conscious Will and Responsibility: A Tribute to Benjamin Libet*. New York: Oxford University Press, Ch. 8.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S. Frith, C. and Passinghuam, R. 2007. Reading hidden intentions in the human brain. *Current Biology* 17: 323–328.

Libet, B. 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences* 8: 529–566.

Northoff, G. 2011. *Neuropsychoanalysis in Practice: Brain, Self and Objects*. New York: Oxford University Press.

2012. Immanuel Kant's mind and the brain's resting state. *Trends in Cognitive Sciences* 16 (July): 356–359.

2014a. *Unlocking the Brain*. Vol. I: *Coding*. Oxford University Press.

2014b. *Unlocking the Brain*. Vol. II: *Consciousness*. Oxford University Press.

Roskies, A. 2008. Neuroimaging and inferential distance. *Neuroethics* 1 (1): 19–30.

Spence, S. 2009. *The Actor's Brain: Exploring the Cognitive Neuroscience of Free Will*. Oxford University Press.

Walter, H. 2001. *Neurophilosophy of Free Will: From Libertarian Illusions to a Concept of Natural Autonomy*, trans. C. Klohr. Cambridge, MA: MIT Press.