MINERVA PSICHIATRICA

VOLUME 55 · N. 1 · MARZO 2014

EDIZIONI · MINERVA · MEDICA

PUBBLICAZIONE PERIODICA TRIMESTRALE - POSTE ITALIANE S.P.A. - SPED. IN A. P. D.L. 353/2003 (CONV. IN L. 27/02/2004 N° 46) ART. 1, COMMA 1, DCB/CN - ISSN 0374-9320 TAXE PERÇUE

MINERVA PSICHIATR 2014;55:1-15

Localization *versus* holism and intrinsic *versus* extrinsic views of the brain: a neurophilosophical approach

G. NORTHOFF

Mind, Brain Imaging and Neuroethics Research Unit
University of Ottawa
Institute of Mental Health Research
Royal Ottawa Mental Health Centre
Ottawa, ON, Canada

The brain can now be visualized in novel ways given the recent emergence of functional imaging techniques like functional magnetic resonance imaging. This revives old question from neuroscience and neurology that have focused on localizing mental states and their contents in particular regions or networks of the brain. Alternative to such localization, a holistic view of brain function where the whole brain participates in processing mental contents has been suggested. Therefore, I here discuss the alternative between localization and holism. In addition, the brain has often been considered to be mere reflex apparatus thus reacting merely mechanical to extrinsic stimuli – neural activity is consequently purely extrinsic. This has been challenged by the recent discovery of intrinsic that is internally generated neural activity in the brain. Therefore, I here contrast extrinsic and intrinsic view of the brain and, at the end, I draw an analogy between the philosopher's Immanuel Kant's view of the mind and an intrinsic view of the brain.

KEY WORDS: Brain - Holistic health - Psychology.

There has been growing development at the interface between neuroscience and philosophy which target brain and mind thus entailing some kind of dualism. There is a rising sense that the dualism between mind and brain needs to be overcome in order to get a grip on the mind and its various mental features in the same way that we can understand the brain only when considering some of mind's mental and psychological features including psychological mechanisms.

While there is much dialog and fruitful exchange between neuroscience and psychology the methodological framework for such convergence remains less clear. One of the main obstacles in linking brain and psyche consists in his inability to locate psychodynamic mechanisms in specific regions or networks of the brain. Psychological contents self, defense mechanisms, memory, unconscious, etc. refer to certain psychological mechanisms that are literally dynamic by their very nature. This contrasts with their localization in specific regions or networks of the brain, that are by definition well defined and hence static rather than dynamic. This amounts to a discrepancy between (psychological) mechanisms and (neuronal) region which methodologically can be described by region- *versus* mechanism-based approach.

Corresponding author: G. Northoff, Mind, Brain Imaging and Neuroethics Research Unit, University of Ottawa, Institute of Mental Health Research, Royal Ottawa Mental Health Centre, 1145 Carling Avenue, Ottawa, ON, Canada.
E-mail: georg.northoff@theroyal.ca

How can we define the concepts of region- *versus* mechanism-based approach? A region-based approach aims to associate a particular psychological content with the neuronal activity in a particular region (like the subgenual cingulate) or network (like DMN). Such region-based approach may be contrasted by a mechanism-based approach that focuses more on linking psychological contents to particular neuronal mechanisms, as it is, for instance, memories are linked to the hippocampus.

Are region- and mechanism-based approach opposite and mutually exclusive? No, they must be regarded rather as complementary and as extremes on a continuum. Ideally, one may want to combine. The concept of dynamic localization that relies on the Russian psychologist AR. Luria opened an initial conceptual bridge between both approaches. How though we really practice rather than merely postulating such combination of region- and mechanism-based approaches? For that we may want to go briefly back into the history of neuroscience and plunge into the debate of localizationism *versus* holism.

The question for localizationism versus holism is closely related of the view of the brain. The brain is often viewed are pure mechanical device that possesses no intrinsic activity that is internally generated neural activity but only neural activity generated by external stimuli that is, extrinsic activity. This is not true, however, given past and recent findings of strong evidence for intrinsic activity in the brain. I consequently discuss in the second part of my paper the difference between what I describe as intrinsic *versus* extrinsic views of the brain as second major topic in any future behavioral neurology and neuropsychiatry.

Neurohistorical remark Ia: "Localizationism" in past and present neuroscience

One of the main methodological approaches in neuroscience at the beginning of the twentieth century was the investigation of patients with brain lesions. These patients could reveal how their higher-order cognitive functions like consciousness, memory, attention, learning, and so on, were affected by lesions in particular regions.

This was the way that early neurologist Paul Broca found out about a specific region in the brain being in charge of comprehending language — the Broca region. He observed that patients with a lesion in the left lateral prefrontal cortex showed major deficits in uttering words and language, a so-called "aphasia". From his clinical observations Broca inferred that this region must be in charge of producing words, thus localizing language in the Broca area, as it is called these days.

Observation of patients with lesions and their corresponding mental disturbances has since been a major tool of insight into the function of the brain. From the exact localization of the lesion and the corresponding mental disturbances, one may infer which region in the brain mediates the respectively underlying higher-order cognitive function. Many other higher-order cognitive functions, including consciousness and self, are currently investigated in this way in neurological patients who suffer from specifically localized lesions in the brain.[1] This entails what I describe as a "localization-based approach" to the brain.

The concept of the "localization-based approach" can be defined in two ways. First, it implies the neuropsychological assumption that a particular function can be related to the neural activity in a specific brain region, meaning that the former can be localized precisely in the latter. This is a neuronal (or better, neuropsychological) meaning of the concept of "localization-based approach" that pertains to a hypothesis about how the brain's regions are related to psychological functions.

In addition to such neuropsychological meaning, the concept of the "localization-based approach" can also refer to an investigator's particular methodological strategy for approaching the brain. The brain here is approached in terms of regions rather than

in terms of, say, processes or codes (see Introduction for such code-based approach to the brain). The methodological approach to the brain in terms of regions is not restricted to the investigation of patients with local brain lesions. It may also extend to the healthy subjects, such as, for instance, those investigated in functional magnetic resonance imaging (fMRI). The use of techniques like fMRI is indeed guided by the search for the localization of particular functions in specific regions of the brain, which it therefore approaches in terms of regions (as distinguished from processes or codes).

Finally, the search for localization of higher-order cognitive functions in patients with brain lesions and functional brain imaging converges with the assumption of modules in cognitive psychology. Cognitive psychology proposed specific functional unities that are in charge of processing and operating such specific cognitive content as attentional content, working memory content, conscious content, self-specific content, and so on. When cognitive psychology entered neuroscience and they were amalgamated into "cognitive neuroscience," the concept of modules was combined with the concept of localization in the brain.[2]

What were described as "modules" in cognitive psychology could then be easily transferred to the brain and more specifically to particular brain regions and their connections. Hence, the localization-based view of brain function seems to be intimately coupled with the module-based view of psychological functioning. This resulted in the assumption of the localization of specific cognitive modules in particular regions (or networks of regions) in the brain.

This is still the implicit or explicit presupposition in current neuroscience and especially in cognitive neuroscience,[3] which is often extended to the more recent branches of affective and social neuroscience: "I take the modular organization of many brain systems as a well-established fact, and discuss only how far fMRI can go in revealing the neuronal mechanisms of behavior by mapping different systems modules and their dynamic interrelationships".[3]

Neurohistorical remark Ib: holism in past and present neuroscience

However, nothing in the science of the brain goes without the opposite suggestion. A strictly localization-based approach was put into doubt early on by another neurologist, Hughlin Jackson, who suggested a more complex and systematic neural organization with multiple interdependencies between different regions. This paved the way for a more holistic view of brain function, one that relates higher-order cognitive functions to the neural operations in the whole brain and its multiple regions.

Interestingly, Sigmund Freud, the founder of psychoanalysis, who initially was a neuroanatomist, also rejected a localization-based approach to the brain. His reason was that more complex psychological disorders, like hysteria or depression, could not be confined to alterations in specific brain regions. He instead regarded these disorders as more complex systems disorders where the organization of the "psychic apparatus," as he called it, is abnormal, which is manifested throughout the whole brain and its different regions. One may therefore consider Freud a forerunner of a more holistic view of brain function.[4, 5]

Later, neuroscientist Karl Lashley [6] observed in his postmortem dissections that the extent of a brain lesion predicts the degree to which higher-order cognitive functions and mental states are disturbed. This let him develop what he called the "Law of equipotentiality" and the "Law of mass action". Both laws describe the distribution of neural processing across the whole brain during higher-order cognitive functions like consciousness and memory. Different regions were proposed to contribute equally to the generation of complex functions that therefore must be considered the result of "mass action" in the brain. This means that higher-order cognitive functions like memory and consciousness were proposed to result from the neural processing throughout the whole brain, rather than being localized in particular regions or modules within the brain.

Analogous observations were made by Russian neuropsychologist A. R. Lurija.[7] Based on his lesion patients, he suggested that one region in the brain can be involved in various higher-order cognitive functions. Conversely, he proposed that higher-order cognitive functions are mediated not only by one or two regions but by various regions in the brain. Most important, the same higher-order cognitive function may even recruit different regions in different instances, depending on the respective psychological and neuronal contexts. There is thus what Lurija described as "dynamic localization". This led Lurija to formulate his hypothesis of functional systems as the operating systems of the brain that describe the actual constellation of different regions that mediate a particular function:

"According to this view a function is, in fact, a functional system […] directed towards the performance of a particular biological task and consisted of a group of interconnected acts that produce the corresponding biological effect. The most significant feature of a functional system is that, as a rule, it is based on a complex dynamic "constellation" of connections, situated at different levels of the nervous system, that in the performance of the adaptive task, may be changed with the task itself may be unchanged".[7]

How about holism in the neuroscience of our days? The earlier-described metabolic approach to the brain by Shulman [2] presupposes a more holistic approach to the brain. By considering the global metabolic-energetic supply and distribution to the brain as a whole as central for any subsequent neural activity, a holistic, and thus global, component is introduced.

Such a more-holistic view is also promoted in parts of functional brain imaging that focus much more on neural networks spanning across different regions rather than on single regions. This is especially apparent in the functional brain imaging of the resting-state activity.[8, 9] However, as we will see further down, even the characterization of the brain by different networks may still presuppose too localizationism.

Finally, the holistic view of the brain also surfaces in the current debate about consciousness in neuroscience. A global workspace of neural activity and information spread are often proposed as central in constituting consciousness; since such a global workspace allows for global extension and distribution, it implies the involvement of different regions and networks throughout the whole brain.[9]

Neurohistorical remark Ic: problems of localizationism in present neuroscience

What is the standing of such a holistic view of brain function these days? The introduction of functional brain imaging has shifted the pendulum back again toward the localization-based view with the assignment of specific regions or networks to particular functions like attention, working memory, and so on.[2]

In addition to the various regions and neural networks supposedly serving specific psychological functions, a network particularly involved in mediating resting-state activity, the default-mode network (DMN), has been distinguished in regional and connectional terms. The DMN seems often (though implicitly) to be regarded as the module for the resting state that therefore stands side by side with other networks that function as modules for specific functions such as, for instance, executive functions or salience.[10]

However, recent imaging studies shed some doubt on the proclaimed localization of specific psychological regions in particular regions or neural networks. The various regions of the DMN, like the anterior and posterior cingulate cortex and the medial prefrontal and parietal cortex, are supposed to serve psychological and mental activity, specifically in the resting state. The same regions are also recruited during a variety of psychological tasks or functions, including contextual association, navigation and spatial processing, episodic memory, decision making, execution errors, self-relat-

ed processing, mind-reading, emotional processing, and social interaction.[11-13]

This sheds some doubt on the regional or network specificity of the DMN; more specifically, on its specific association with particular psychological functions during either resting-state activity or stimulus-induced activity. Conversely, these observations also argue against region-specific (or network-specific) localization of the various functions themselves, which seem to recruit more or less the same regions and networks.

This situation with the recruitment of the same regions and network by different functions is not peculiar to the DMN. The same pattern can be observed in the case of another neural network that includes the bilateral anterior insula, the dorsal anterior cingulate cortex, and the thalamus as its core regions (these regions are also subsumed under what is described as the "salience network" [10]). These regions are active during functions as diverse as interoceptive awareness,[14, 15] empathy,[16] anticipation of emotions,[17] and aversion.[18] The list of regions that are recruited by different functions can easily be extended.

In sum, the observation of the same region and network mediating a variety of different functions sheds some doubt upon the localization-based approach and its attempts to establish a specific one-to-one relationship between regions/networks and functions.

Does this mean that we have to revert to a more holistic view of the brain and its different regions? Based on their data, some neuroscientists — doing either lesion-based studies [1] or functional imaging using electroencephalography (EEG);[19] positron emission tomography (PET);[2] or functional magnetic resonance imaging (fMRI);[20] — do indeed advocate a more holistic view of brain function. This is further corroborated by neuroanatomy, which considers single regions as hubs or nodes within the neural network of the whole brain rather than as centers or modules by themselves.[21, 22]

Where does this leave us? Do we have to follow the swings between localizationism and holism? My aim in the following discussion is to show how both are very compatible and complementary, rather than being contradictory. And it this complementarity which may be essential for understanding how the brain's neuronal mechanisms are linked to and, even stronger, make necessary the kind of psychodynamic mechanisms so well described in the various contributions in this book.

Neurotheoretical remark Ia: localization and sparseness

While the association of a specific region or network with a specific psychological function must be considered doubtful, the data nevertheless show that only a certain set of regions is recruited during the various tasks or functions. Multiple functions seem to recruit the same set of regions or network entailing a many/multiple-to-one/few relationship between functions and regions. The function-region relationship thus seems to obey the rules of sparseness, with sparse representation of the multiple functions in a few regions/networks of the brain. I consequently hypothesize sparseness rather than localization to operate and determine the function–region relationship.

More specifically, I propose that what is considered localization of a particular function in a specific region reflects the sparse number of actually activated regions when compared to the total number of regions that could possibly be recruited. The fact that the other regions are not activated does not mean, however, that they do not participate in generating the function in question.

The inactive regions may nevertheless have an important role in that their baseline — that is, resting-state activity — may serve to generate and amplify neural differences (presupposing difference-based coding on a regional level [8]). These neural differences may in turn allow the brain to condense and sparsen neural activity in one or a few subsequent regions, yielding those regions that we observe to be activated. Accordingly, sparse coding on a regional level seems

to be nicely compatible with the proposed localization of particular functions in specific regions.

How does the assumption of such sparse coding stand in relation to the localization approach? To equate sparse coding with localization is to confuse the underlying processes and their resulting outcomes. The localization-based approach focuses on the outcome while neglecting the process itself; that is, how the apparent localization of a function in a particular region is generated. Instead of considering the process of generating regional localization, the localization-based approach takes the localization of a particular function in a specific region for granted. And it considers the psychological function to be intrinsic or innate to the region itself without further questioning the underlying processes how that function is generated by the region's neural activity.

Such a localization-based approach is, however, to be distinguished from the approach sparseness-based approach takes to the question of localization. Here the focus shifts from the outcome, the observation of a regional localization, to the processes; that is, the rules and principles that generate what we observe and propose as the specific linkage between function and region. This means that the psychodynamic mechanisms in question like repression or ego is necessarily implied by the specific neuronal mechanism that characterizes a particular region (like subgenual) or network (like DMN) in its relationship to others.

More specifically, one can then no longer say that the psychodynamic mechanism is localized in the respective region or network. Instead, the observed region or network is just a convergence of that region's/network's relationship to others as reflected in the postulated neuronal mechanism that operates across the boundaries of different regions and networks. What is needed would then be the investigation of the neuronal mechanisms by means of which neuronal mechanism neuronal activity across the different regions and networks becomes focused and crystallized and thus sparsedin the subgenual cingulate.

Does such sparseness on the neuronal level may correspond to sparseness on the psychic level such that repression may be the result of prior sparsening of psychic energy and resources? This question has both methodological and empirical-conceptual implications. Methodologically, if so, psychological mechanisms and contents may help us in guiding and understanding how the brain organizes and structures its own neuronal activity such that it results in the kinds of regions and networks we observe as the very output of prior and preceding processes like sparsening that hitherto remain largely unknown.

While empirical-conceptually, we may then assume close if not intrinsic and necessary rather than merely extrinsic and contingent linkage between neuronal and psychological mechanisms: the latter do not simply supervene upon the former in a contingent way but are instead are already predisposed in a necessary way by the brain's structuring and organizing of its neuronal activity by neuronal mechanisms like sparsening that operate across different regions and networks. The shift from regions and networks to neuronal mechanisms like sparsening that may potentially underlie the activity distribution into different regions and networks may then provide a true that is intrinsic link between psyche and brain.

Neurotheoretical remark Ib: distinction between "activated" and "active" regions

How can such sparsening of neuronal activity be realized and implemented? Briefly, I propose that the activation of a specific region yields from the computing and comparing of neural differences stemming from other regions (see Northoff a for details of such difference-based coding [8]). These regions, which serve to yield and amplify neural differences, may by themselves either be activated or non-activated. This means that even non-activated or non-recruited regions participate in generating neural differences.

Conceptually, one may therefore want to distinguish between "activated" and "active" regions. "Activated regions" are those regions that show neural activity changes in response to the task we apply. We as observers propose these regions to be recruited by the function in question and are consequently inclined to localize the latter in the former.

This, however, neglects what I describe as "active regions" that do not show changes in their activity level in response to the task. These regions may nevertheless participate in generating the neural activity changes of the activated regions, more specifically in generating and amplifying neural differences (what can be described as an "amplification hypothesis"; see Northoff 2013a). They are thus "active" but not "activated". This, however, makes localization of the function in the activated regions impossible, since that would neglect the role of the active regions in generating the neural activity changes in the activated region.

The generation and amplification of neural differences is coupled to the condensation of neural activity (what can be described as a "condensation hypothesis"; Northoff 2013a). Rather than each of the original lower order sensory regions' activating a separate higher order cognitive region, the former's neural activity converges in one common region, to which we then attribute localization. This, however, is a false-negative inference that focuses only on the outcome of localization in the higher-order cognitive region, while neglecting its underlying processes in which lower-order sensory regions participate.

More specifically, the outcome of sparse coding does indeed pertain to one particular region, the "activated" region or network as distinguished from all "non-activated" regions/networks. However, the underlying process involves "active" regions/networks (as distinguished from non-active regions/networks) that are essential in yielding and amplifying neural differences.

This means that the function in question cannot be localized exclusively and completely in the "activated" region/network itself. Instead, the function may be associated with both "activated" and "active" regions/networks as distinguished from "non-activated" and "non-active" ones. Accordingly, the regions/networks remaining silent in response to our task, that is, "non-activated," may nevertheless be "active" (rather than "non-active") and may therefore have an important role in processing the function in question (Figure 1; see also Hayes *et al.*[23] for an example of where the density of GABA-A receptors [PET] in ventromedial prefrontal cortex, a non-activated region during an aversive task, modulates the degree of signal changes [fMRI] in an activated region, the sensorimotor cortex; see also Gonzales-Castillo *et al.*[24] for the support of active but non-activated regions).

In sum, one may well propose localization of neural activity in specific "activated" regions/networks during particular function. This, as demonstrated, is the outcome of the processes guiding sparse coding on a regional level. However, to infer from such localization (or better, condensation) of neural activity to the localization of the function in question in that particular region/network is to confuse outcome and processes. The function in question must also be associated with regions (and networks) other than the "activated" ones like those that I here described as "active"; that is, actively involved in amplifying neural differences.

Neurotheoretical remark Ic: complementarity between holism and localizationism

We are confronted with two apparently contradicting observations. On one hand, many regions, and ultimately, the whole brain, seem to be implicated in the neural processing of various psychological functions (see earlier). This suggests holism holds true on a psychological level. On the other hand, there is regional sparseness in that different psychological functions seem to recruit similar but at least strongly overlapping regions and networks. This observation, however, contradicts holism and

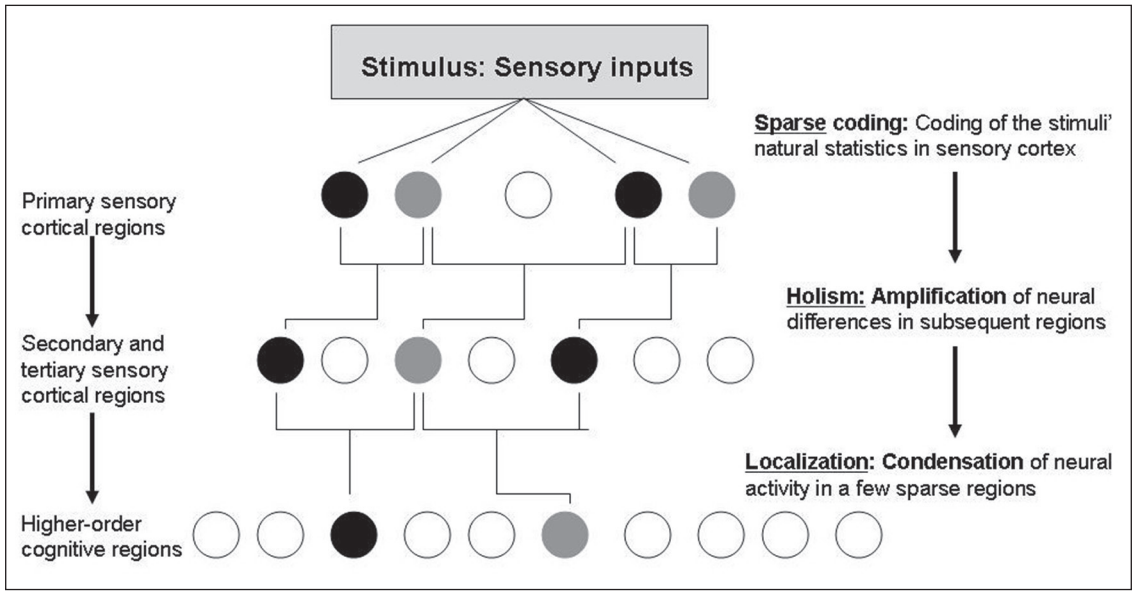

Figure 1.—Complementarity between holism and localizationism. Black: Activated/recruited regions; Gray: Non-activated but active regions participating in yielding neural differences; White: Non-activated and non-active regions. The figure depicts the different stages of neural processing. The stimulus is encoded into the sensory cortex's neural activity a sparse way; *i.e.*, as based on its natural statistics as its statistical frequency distribution across different discrete points in time and space. This is possible only if we presuppose difference, rather than stimulus-based coding (upper part). Even if regions are not activated or recruited by themselves, they may still participate in constituting neural differences; they are thus "non-activated" but nevertheless active. The initial neural differences in primary sensory cortex are supposed to be amplified ("amplification hypothesis") in subsequent regions, entailing holistic distribution of the initial neural activity changes across different regions of the brain; i.e., holism (middle part). That in turn makes possible the condensation of neural differences ("condensation hypothesis") in a few subsequent regions that then do show up as "activated" regions (lower part). These different stages of neural processing across the different regions of the brain are well reflected in changing ratios between "activated" (or recruited) regions, "non-activated" but "active" regions, and "non-activated and non-active" regions.

would rather be compatible with localizationism holding true on a neuronal level.

How can we reconcile the contradictory assumptions of localizationism on the neuronal level and holism on the psychological level? The need to reconcile localizationism and holism was already recognized by K. Lashley, as is apparent in the following passage:

"The chief advantage of the strict theories of localization has been their definiteness and comprehensibility. Those of us who have felt the inadequacy of such theories have had to fall back upon expressions like mass action, stress patterns, dynamic effects, melodies of movement, vigilance or nervous energy; all metaphorical and highly unproductive of experimental problems. Yet the facts demand something of this sort. The evidence seems conclusive that in various cortical functions there is every degree of specialization from a limited point-to-point correspondence of cells to a condition of absolute non-specificity. Not only is there diversity in the modes of action of different parts of the cortex but a single area, highly specialized and differentiated for one activity may be wholly undifferentiated for another in which it also participates. We have not a choice between a theory of localization and a theory of decentralization, but must develop a wider view which recognizes the importance and interdependence of both modes of integration".[6]

I hypothesize that we need to set the alternative of localizationism *versus* holism into the context of sparse coding and difference-based coding in order to reconcile both. There is holism on the process level that is, on the level of neuronal mechanisms that operate across regions and networks. As described earlier, even presumably silent, that is, "non-activated" but "active", regions/net-

works are nevertheless actively participating in generating and amplifying neural differences, thus allowing for difference-based coding. Such difference-based coding is in turn central in condensing and thus sparsening neural activity in a particular region, the "activated" or recruited region, as the manifestation of sparse coding on a regional level.

Many regions, if not (indirectly via the constitution of differences) the whole brain, actively participate in constituting neural differences. One may consequently propose holism on the level of neuronal processes, whereas the very same neuronal processes, operating throughout different regions, allow and, even stronger, predispose the temporal and spatial sparsening of subsequent neural activity changes in a few highly localized regions as their outcome. The outcome, that is, the changes in the neural activity in a few localized regions, may thus be more localized when compared to the rather holistically operating processes.

What does this entail for the relationship between localizationism and holism? This means that the concepts of localizationism and holism are not opposite and contradictory to each other but rather mutually dependent on each other: As there would be no outcome without a preceding process, localizationism would remain impossible without holism. Even stronger, the more holistically processes that allow for the amplification of neural differences throughout the whole brain make a more localized outcome, that is, spatial and temporal sparsening of neural activity and the number of "activated" regions, almost necessary. Accordingly, localizationism and holism are bound together as tightly as process and outcome; they remain consequently as inseparable and complementary as yin and yang in the Chinese tradition.

Neurotheoreticla remark Id: mechanisms with mechanisms and outcomes with outcomes

What does this now imply for the relationship between mind and brain? The localization of neuronal activity in particular regions or networks is just the outcome or final result of a long chain of yet unclear neuronal processes and mechanisms like spatial and temporal sparsening that operate across different regions and networks.

How does that stand in relation to mental and psychological mechanisms? Psychological mechanisms like repression refer to processes in the mind by means of which a particular outcome or result is achieved, the disappearance of a particular psychological or mental content. The disappearance of a mental or psychological content as the outcome or final result of an underlying psychological mechanisms corresponds conceptually on the neuronal side to the localization of neural activity in a particular region or network that, analogously, is the outcome or final result of equally complex underlying neuronal mechanisms (like temporal and spatial sparsening).

Accordingly, to now locate a psychological contentsin the neural activity of a particular region or network is to confuse a particular mechanism (like sparsening on the neuronal level) with its outcome or result (like the disappearance of a particular content or the activity of a specific region or network).

In the same way we distinguish between the car ride itself and the city we arrive at the end of our 12 hour car ride, we need to distinguish between mechanisms as the ride that lead to a certain outcome from the actual outcome itself. Only if we link mechanism to mechanism and outcome to outcome in our attempt to bridge the gap between psyche and brain, we will be able to understand why and how our brain and its resting state and stimulus-induced activity that signify yet unknown neuronal mechanisms are so intimately that is, intrinsically and necessary rather than extrinsically and contingently, linked to psychological contents and mechanisms. One may then want to speak of neuropsychological mechanisms rather than either psychological or neuronal mechanisms.

Finally, the difference between region- and mechanism-based approaches will

then no longer appear as contradictory but complementary entailing the crystallization of neural activity in particular regions or networks. This will further tighten the link between psyche and brain in such way that in 20-30 years we will hopefully ask ourselves why we ever considered it even a problem to link neuronal and psychological mechanism in very much the same way we nowadays do not see any problem in the fact that the 12 hour car ride lead us to the city of our dreams.

View of the brain Ia: intrinsic *versus* extrinsic view of the brain – History of neuroscience

So far we have discussed how the brain and its various regions and networks are related to the different functions it generates. This leaves open how the brain operates and processes the various kinds of inputs it receives from the body and the environment. Any empirical characterization of the brain as a whole must, therefore, ask questions about the brain's mode of operation. What is the brain and how does it operate? This was the subject of controversial discussion at the beginning of the twentieth century. One view, favored by the British neurologist Sir Charles Sherrington (1857-1952), proposed the brain and the spinal cord as primarily reflexive.

Reflexive describes that the brain reacts in predefined and automatic ways to stimuli. This means that the stimuli from the outside of the brain, originating extrinsically in either the body or environment, are proposed to determine completely and exclusively the subsequent neural activity. The resulting stimulus-induced activity, and more generally any neural activity, in the brain can thus be traced back to the extrinsic stimuli. This is what I describe as the "extrinsic view" of the brain (see upper part in Figure 2).

An alternative view, however, was suggested by one of Sherrington's students, Thomas Graham Brown. In contrast to his teacher, he suggested that the brain's activity, that is, in the spinal cord and the brain stem, is not primarily driven by extrinsic stimuli from the outside of the brain. Instead, the spinal cord and the brain stem show spontaneous activity originating intrinsically, or within themselves. Other subsequent neuroscientists like Karl Lashley, Kurt Goldstein, and Wolfgang Koehler followed Brown's line of thought and supported the hypothesis of intrinsic activity. This is called the "intrinsic view" of the brain (see lower part in Figure 2).

The assumption of intrinsic activity generated inside the brain itself has major implications for how we conceive stimulus-induced activity. What we as observers describe as stimulus-induced activity, and usually associate with the stimulus itself, must then be regarded as the hybrid result of a specific interaction between the brain's intrinsic activity and the extrinsic stimulus. Stimulus-induced activity and any neural activity in the brain must be traced back to a double input that originates in both the brain's intrinsic activity and the body and the environment's extrinsic stimuli.

View of the brain Ib: intrinsic *versus* extrinsic view of the brain – Present findings

What does current research tell us about the brain? The dichotomy between intrinsic and extrinsic views of the brain is still as controversial. It has especially resurfaced in functional brain imaging. Let's start with the extrinsic view. Many levels of neuroscience, ranging from cellular to regional to behavioral levels of the brain, rely on experimental application of specific stimuli and tasks to probe neural activity. By comparing different stimuli and tasks, the resulting differences in neural activity are associated with the respective stimuli or tasks. This means that the experimental requirements may predispose and pull us toward an extrinsic view. The extrinsic view has been most predominant in behaviourism. Behaviorism is the characterization of mental states by mere input-output, *e.g.*, stimulus-response

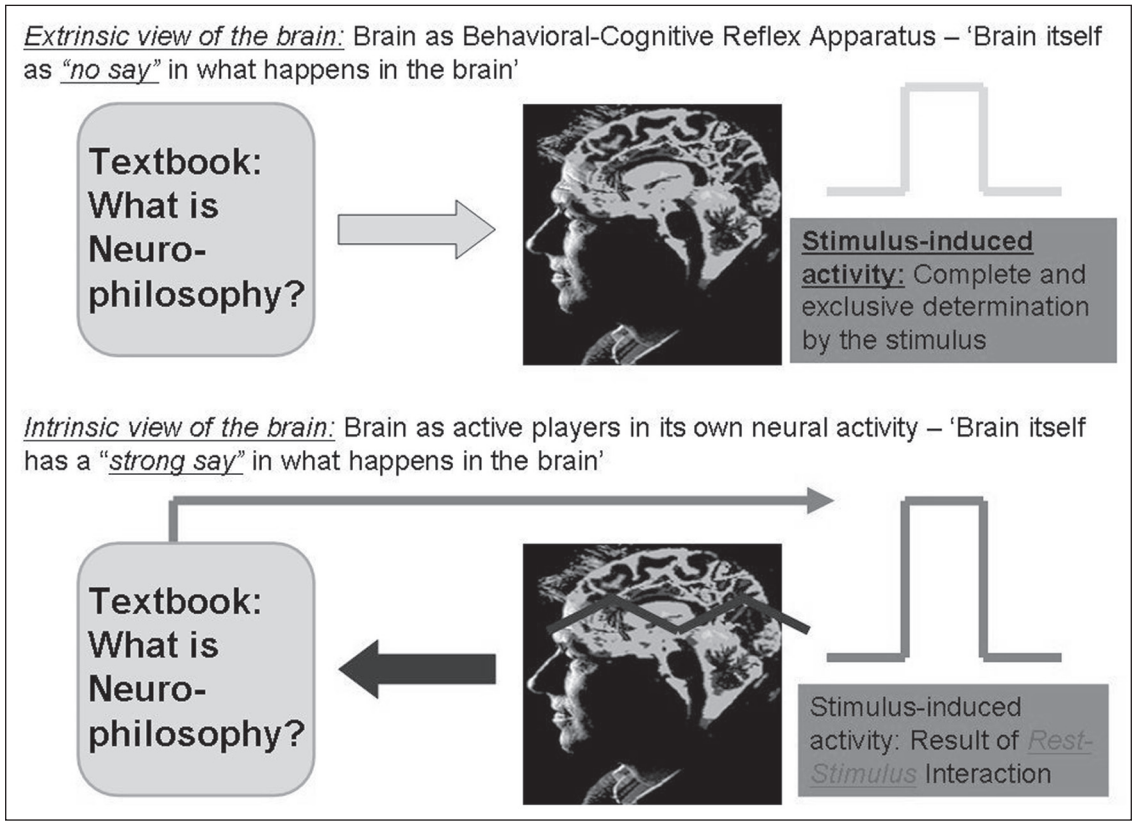

Figure 2.—Extrinsic versus intrinsic view of the brain. This figure illustrates different views of the brain, extrinsic (A) and intrinsic (B). A) In the extrinsic view of the brain, the neural activity in the brain is exclusively related to the external stimulus as here indicated on the left by the cover of a textbook on neurophilosophy. This in turn induces what is called stimulus-induced activity as indicated on the right. That means that the brain itself does not make an active contribution to the neural activity the stimulus generates inside the brain and, metaphorically speaking, has no say in what happens in the brain; B) in the intrinsic view of the brain, the stimulus-induced activity is considered to result from an interaction between the effects of the stimuli in the brain and the intrinsic activity as contributed by the brain itself. This means that the brain provides an active contribution to the resulting stimulus-induced activity, namely its own intrinsic activity and how it can interact with the stimulus. Hence, metaphorically speaking, the brain has a strong say in what happens in the brain as induced by the stimulus.

relations, which finds its continuation in contemporary cognitive neuroscience.

However, the extrinsic view of the brain has most recently been challenged again on several grounds. Even in the resting state, that is, in the absence of any kind of particular extrinsic stimuli from either the body or the environment, the brain shows a high degree of metabolic consumption, for instance, about 20% of the body's overall energy budget (and oxygen fraction).

Using functional imaging, this high metabolism has been observed especially in a particular set of regions, the default-mode network (DMN), which includes various anterior and posterior cortical midline structures, as well as the bilateral posterior parietal cortex. The high degree of metabolism is indicative of continuously ongoing high levels of neural activity even in the absence of (specific) extrinsic stimuli — that is, in the resting state of the DMN.

However, other regions outside the DMN also show spontaneous neural activity independent of any extrinsic stimuli. This has been, for instance, demonstrated in the auditory and visual cortices, the thalamus, the hippocampus, the olfactory cortex, the cortical midline regions, the prefrontal cortex, the motor cortex, and other subcortical

regions like the brain stem and the midbrain. The metabolic and neuronal signs of intrinsic activity are further complemented by behavioral evidence. Such behavioral evidence can for instance be observed by what is described as seeking: seeking refers to the continuous search for reward in the environment even in the absence of extrinsic stimuli. As a result, seeking reflects continuous behavioral activation, *e.g.*, arousal.

View of the brain Ic: intrinsic *versus* extrinsic view of the brain – Interaction

Which view holds — the intrinsic or the extrinsic one? Rather than cashing out one view at the expense of the other, the brain itself may force us to go beyond and to reconcile both views. Any given neural activity in the brain may be proposed to result from the interaction between the brain's intrinsic activity and the extrinsic stimuli from either the body, *e.g.*, interoceptive stimuli, or the environment, *e.g.*, exteroceptive stimuli.

Why? Even in an apparent resting state like sleep, the seemingly intrinsic activity of the brain is nevertheless exposed to continuous extrinsic, *e.g.*, intero- and exteroceptive input from the body and the environment. For instance, the continuous action of heart sends interoceptive stimuli to the brain as do the continuous tactile, auditory, olfactory, and gustatory stimuli from the environment during sleep. Conversely, any extrinsic stimulus first encounters the brain's intrinsic activity before it can be processed at all and consecutively associated with sensorimotor, affective, cognitive, and social functions, as well as their respective neural systems in the brain. Hence, there is no pure intrinsic activity, just as there is no pure extrinsic activity in the brain.

Therefore, rather than opposing intrinsic and extrinsic views, we may need to investigate how intrinsic activity and extrinsic stimuli interact with each other in order to understand the brain's neural activity. The neural activity we observe in the brain in response to specific extrinsic stimuli from either body or environment may then be regarded as the result of prior modulation of the brain's intrinsic activity by the extrinsic stimulus.

View of the brain IIa: intrinsic activity – Purpose of the brain's intrinsic activity

What does the brain do with this huge amount of energy in the resting state? The brain uses part of this intrinsic energy to process extrinsic stimuli. To do this, it spends only a tiny fraction, around 2-10%, of its total energy budget. What happens to the rest? We do not know yet. One may consider the brain's high resting state activity as mere noise in the background of stimulus-induced activity. The resting state activity may then be regarded just as a bad side effect of the stimulus-induced activity, which is the main focus of the brain. But why then does the brain waste so much energy and effort for mere noise?

The brain invests the largest part of its energy into its resting state activity. Why is the resting state activity so important for the brain and apparently "much more" than mere noise? The resting state activity may be the "tool" by means of which the brain can affect the stimulus and how it is processed in the brain. This means that there must be some kind of interaction between the brain's resting state activity and the stimulus — this is called rest-stimulus interaction.

Rest-stimulus interaction describes the brain's resting state activity as affecting and ingraining itself upon the neural activity changes induced by the extrinsic stimuli the brain encounters from the environment. Recent studies demonstrated that this is indeed the case when the resting state activity level in the visual cortex predicts the subsequent stimulus-induced activity and its associated behavioural and mental states.

Despite the major advances in recent neuroscience, many issues about the brain's resting state activity remain unresolved these days. What is clear though is that we need a

view of the brain different from the extrinsic and cognitive view as tacitly presupposed in current cognitive neuroscience. Cognitive neuroscience focuses almost exclusively on stimulus-induced or task-related activity. Since the respective stimuli or tasks come outside the brain, their related neural activities may be described as extrinsic.

This contrasts with the resting state activity that originates from the inside of the brain and may therefore be designated as intrinsic. A shift in focus from stimulus-induced to resting state activity consequently entails a shift from an extrinsic to an intrinsic view of the brain. Interestingly, the quest for an intrinsic view of the brain is not specific to our time, but surfaced in neuroscience at the beginning of the 20th century. The implications of the intrinsic view of the brain in understanding the mind remain unclear at this point, and are currently hot research topics in neuroscience. The empirical details of the brain's intrinsic activity and its functional implications will therefore be discussed in the subsequent sections on consciousness and the self.

How does the intrinsic resting state activity of the brain interact with the extrinsic stimuli from the outside world? The relevance of such rest-stimulus interaction is supported by recent findings that show that the level of prestimulus resting-state activity predicts the neural, phenomenal, and behavioural effects of subsequent stimuli. What remains unclear, however, are the exact neuronal features of the resting state itself that make such rest-stimulus interaction possible. These neuronal features must be intrinsic to the resting state itself, while at the same time they must also be able to create the tendency (*i.e.*, neural predisposition) to associate stimulus-induced activity with consciousness and self.

Hence, in order to better understand observations during rest-stimulus interaction, we may need to achieve a better understanding of the resting state's intrinsic features. Additionally, we must learn how they predispose rest-stimulus interaction in such a way that the stimulus becomes associated with consciousness and self. In order to do this, we may need to develop an intrinsic-extrinsic interaction model with regard to the brain.

View of the brain IIb: intrinsic activity Kant's mind and the brain's intrinsic activity

In order to develop an intrinsic-extrinsic interaction model, we may want to venture on a brief excursion in the history of philosophy. The French philosopher Rene Descartes assumed mental properties intrinsic to the mind to be distinct from the physical features of body and brain. The Scottish philosopher David Hume, who opposed such intrinsic mental properties, disagreed. Instead, Hume advocated an extrinsic view of the mind, believing that mental activity can be entirely traced back to the extrinsic features of stimuli in the world.

His German successor, Immanuel Kant (1998), combined both intrinsic and extrinsic views of the mind: he claimed that consciousness and self must be considered a hybrid of processes resulting from an interaction between the mind's intrinsic features, and the world's extrinsic stimuli. In order to reveal the nature of such intrinsic-extrinsic interactions, Kant attributed various faculties (*i.e.*, intrinsic features) to the mind, primarily described in his *Critique of Pure Reason*.[25]

The mind's intrinsic features included unity of consciousness, self as "I think", and various templates of spatiotemporal continuity (which were subsumed under the umbrella term "categories"). According to Kant, the mind uses its intrinsic features to structure and organise the effects of extrinsic stimuli. This, in turn, allows the latter to become associated with consciousness, self, and spatiotemporal continuity. Hence, consciousness, self, and spatiotemporal continuity are based on the interaction between the mind's intrinsic features and the environment's extrinsic stimuli.

How is Kant's view of the mind and its intrinsic features related to the brain's intrinsic activity? Kant characterized the mind

by the process of thinking expressed by "I think". In addition, he attributed other intrinsic (*i.e.*, transcendental) features to the mind, like unity, and distinguished them from mere extrinsic (*i.e.*, empirical) stimuli. How can Kant's transcendental view of the mind's intrinsic features inform the neuroscientific investigation of consciousness?

The concept of "I think" entails that any cognition of extrinsic stimuli must be accompanied by the "I" (*i.e.*, the self) and its thinking activity as an intrinsic feature of the mind itself. Why is that necessary? This is where neuroscience can shed light on Kant's thought. If the resting state is indeed organized in a self-specific (or self-perspectival; see Part V for details) way, no extrinsic stimulus can "avoid" the encounter with the resting state, that is, rest-stimulus interaction and its association with the self.

View of the brain IIc: intrinsic activity Kant's consciousness and the brain's consciousness

How can Kant help inform neuroscientific investigation? Kant deemed "I think" to be essential for consciousness: we cannot be conscious without the mind's accompanying "I think". If "I think" is indeed related to resting-state activity, it may help decipher the neuronal features of the resting state and its role in consciousness. Current neuroscientific research focuses mainly on stimulus-induced activity, which is supposed to be sufficient for consciousness, the neural correlate of consciousness (NCC). This, however, neglects one central feature, 'I think' and, in neuronal terms, resting state activity. Resting state activity itself must contain certain features that are central in constituting consciousness.

In the same way that Kant suggested "I think" to be necessary for consciousness, we may assume the resting state to be necessary for associating stimulus-induced activity with consciousness. The resting state may then be regarded as a necessary, non-sufficient condition, a neural predisposition of consciousness (NPC).[8, 9]

What are the features of the resting state that predispose consciousness? Besides "I think", Kant considered "unity" (*i.e.*, transcendental unity) to be an intrinsic feature of the mind. Following Kant, one may assume a particular, but currently unknown, unity of neuronal activity in the resting state to predispose consciousness. Relying on Kant, the British neuroscientist Semir Zeki,[26] assumes such unity to be pre-programmed and central in the neural constitution of visual consciousness. Following Kant, such unity must be described as neurotranscendental, as it must be predisposed by the resting state itself and its specific but unknown spatiotemporal organization. This in turn may, for instance, make the binding and grouping of different stimuli in consciousness as discussed in the binding problem possible.

What can we learn about the brain from Kant's mind? Future work may want to explore the exact neural mechanisms underlying different forms of rest-stimulus interaction. This, however, is possible only if we achieve a better understanding of the neuronal mechanisms underlying the brain's intrinsic activity, its resting state. In order to achieve that, we may draw on Kant's insights about the mind's intrinsic features such as "I think" and unity. This may allow us to better understand how the brain's resting state activity is structured and organized. And, most importantly, how that predisposes certain kinds of rest-stimulus interaction and, as Kant might say, consciousness and self.

References

1. Feinberg TE. From axons to identity: neurological explorations of the nature of the self. New York: WW Norton & Company; 2009.
2. van Eijsden P, Hyder F, Rothman DL, Shulman RG. Neurophysiology of functional imaging. NeuroImage 2009;45:1047-54.
3. Logothetis NK. What we can do and what we cannot do with fMRI. Nature 2008;453:869-78.
4. Northoff G. Neuropsychoanalysis in practice: brain, self and objects. Oxford: Oxford University Press; 2011.
5. Northoff G. Immanuel Kant's mind and the brain's resting state. Trends Cogn Sci 2012;16:356-9.
6. Lashley KS. Persistent problems in the evolution of mind. Q Rev Biol 1949;24:28-42.

7. Luria AR. The working brain: an introduction to neuropsychology. Trans. by Haigh, B. London: Penguin; 1973.
8. Northoff G. Unlocking the brain. Vol. I: Coding. Oxford: Oxford University Press; 2013a.
9. Northoff G. Unlocking the Brain. Vol. II: Consciousness. Oxford: Oxford University Press; 2013b.
10. Menon V. Large-scale brain networks and psychopathology: a unifying triple network model. Trends Cogn Sci 2011;15:483-506.
11. Bar M. The proactive brain: using analogies and associations to generate predictions. Trends Cogn Sci 2007;11:280-9.
12. Bar M. The proactive brain: memory for predictions. Philosophical transactions of the Royal Society of London Series B. Biological Sciences 2009;364:1235-43.
13. Spreng RN, Mar RA, Kim AS. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J Cogn Neurosci 2009;21:489-510.
14. Critchley HD, Wiens S, Rotshtein P, Ohman A, Dolan RJ. Neural systems supporting interoceptive awareness. Nat Neurosci 2004;7:189-95.
15. Wiebking C, Bauer A, de Greck M, Duncan NW, Tempelmann C, Northoff G. Abnormal body perception and neural activity in the insula in depression: an fMRI study of the depressed "material me". World J Biol Psychiatry 2010;11:538-49.
16. Fan Y, Duncan NW, de Greck M, Northoff G. Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. Neurosci Biobehav Rev 2011;35:903-11.
17. Bermpohl F, Walter M, Sajonz B, Lucke C, Hagele C, Sterzer P *et al*. Attentional modulation of emotional stimulus processing in patients with major depression alterations in prefrontal cortical regions. Neuroscience Letters 2009;463:108-13.
18. Hayes DJ, Northoff G. Identifying a network of brain regions involved in aversion-related processing: a cross-species translational investigation. Front Integr Neurosci 2011;5:49.
19. John ER. The sometimes pernicious role of theory in science. Int J Psychophysiol 2006;62:377-83.
20. Northoff G. Neuropsychiatry. An old discipline in a new gestalt bridging biological psychiatry, neuropsychology, and cognitive neurology. Eur Arch Psychiatry Clin Neurosci 2008;258:226-38.
21. Hagmann P, Cammoun L, Gigandet X, Meuli R, Honey CJ, Wedeen VJ *et al*. Mapping the structural core of human cerebral cortex. PLoS biology 2008;6:e159.
22. Sporns O. Networks of the brain. Cambridge, MA: The MIT Press; 2011.
23. Hayes DJ, Duncan NW, Wiebking C, Pietruska K, Qin P, Lang S *et al*. GABAA receptors predict aversion-related brain responses: an fMRI-PET investigation in healthy humans. Neuropsychopharmacology 2013;38:1438-50.
24. Gonzalez-Castillo J, Saad ZS, Handwerker DA, Inati SJ, Brenowitz N, Bandettini PA. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. Proceedings of the National Academy of Sciences of the United States of America 2012;109:5487-92.
25. Kant I, Guyer P, Wood AW. Critique of pure reason. Cambridge: Cambridge University Press; 1998.
26. Zeki S. The disunity of consciousness. Progress in Brain Research 2008;168:11-8.

Conflicts of interest.—The authors certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.