*Sebastjan VÖRÖS and Olga MARKIČ**

PHOTOGRAPH OF THE SOUL: TOWARDS A CRITICAL NEUROSCIENCE

Abstract. *The remarkable development of neuroscience in the past three decades (the so-called "neuroscientific revolution") has had a tremendous impact on our understanding of ourselves and the world. The slow, but persistent spread of neuroscience into humanities, social sciences, and everyday life has prompted several authors to critically examine and reassess some of its far-reaching claims, along with its methods of collecting, organising and interpreting data. It has been increasingly pointed out that there is a profound difference between what neuroscience purports to explain and what it actually does and can explain, and that therefore a better understanding of conceptual, methodological, and metaphysical presuppositions of neuroscience is needed. The goal of this article is to provide a small contribution to this vast critical endeavour, focusing particularly on the prominent modular hypothesis, i.e. the idea that the mind consists of a plethora of different cognitive functions (modules) and that these are somehow instantiated or realised in discrete brain regions. After delineating some of the major shortcomings of the modular thesis, the article goes on to argue that it is essential for neuroscience to become better acquainted with its underlying assumptions, and that a platform for constructive and engaged dialogue with other areas of research is needed.*
Keywords: *critical neuroscience, social neuroscience, modularism, phrenology, epistemology, philosophy of science*

Neurorevolution – Myth or Fiction?

There has been much talk of "neuroscientific revolution" (Lynch, 2009) in the past three decades. Ever since the development of new brain scanning techniques (fMRI, PET, SPECT, etc.)[1] in the 1990's, there has been a

* *Sebastjan Vörös, PhD, Teaching Assistant; Olga Markič, PhD, Professor, Faculty of Arts, University of Ljubljana.*

[1] *fMRI – functional magnetic resonance imaging, PET – positron emission tomography, SPECT – Single-photon emission computed tomography.*

growing conviction among many neuroscientists and philosophers that the only appropriate explanatory framework of mental and social phenomena is that provided by the brain sciences. As Francis Crick put it bluntly, "You, your joys and sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behaviour of a vast assembly of nerve cells and their associated molecules" (Crick, 1994: 3). The prospect of being able to determine the neurobiological underpinnings of various aspects of mental and social life proved attractive to experts from different scientific backgrounds, who believed that the dazzling brain imagery might finally provide the "proper scientific foundations" for their disciplines. The past twenty years have thus witnessed a plethora of neuro-disciplines: from neuroeconomics, neurolaw, neuroethics, neuromarketing, and neuropolitics, to neurohistory, neuraesthetics, neurophilosophy, and even neurotheology (Satel and Lilienfeld, 2013: ix; Tallis, 2012: 60). "Neurotalk" (Illes et al., 2010) has been slowly seeping not only into humanities and social sciences, but also into the crooks and crevices of our everyday life.

But it was not long before several authors (Bennett and Hacker, 2003; Choudhury and Slaby, 2012a; Satel and Lilienfeld, 2013; Tallis, 2012; Uttal, 2001) started voicing their concerns about explanatory strategies promulgated by neuroscience. Some of its far-reaching claims and interpretations, along with its methods of collecting, organising, and interpreting data, were subjected to fierce criticism from philosophical, ethical, sociological, and historical perspectives. There has been talk of "neuromania", "neurohubris", and "neurohype" (Satel and Lilienfeld, 2013: xiv), even of "neuromanic imperialism" (Tallis, 2012: 73), and objections have been raised that there is a profound difference between what neuroscience *purports* to (be able to) explain and what it actually *does* and *can* explain. The problem seems especially pertinent in the case of newly emerging neuro-disciplines that have set out to provide neuroscientific accounts of very complex phenomena, e.g. ethics, politics, religion, etc. It has namely become increasingly obvious that it is far from clear what the colourful brain images *actually* tell us about moral belief, political decision-making and religious experience, and that therefore a better understanding of conceptual, methodological, and metaphysical presuppositions of neuroscience is needed.

For this and other reasons, Choudhury and Slaby (2012b) have questioned the talk of a "neurorevolution", suggesting that "the breathless convictions that within a few years (...) the brain sciences will (...) begin to supersede social, cultural, philosophical, political, literary, or other 'folk' explanations of behavioral phenomena" are exaggerated and speculative. Instead of focusing solely on prophecies made by neuroscientists, scholars should also critically engage with "the assumptions and visions

of neuroscience on which such [future] scenarios are built", as these are equally, if not more important in elucidating the reasons why neuroscience has gained such widespread recognition in the academic circles and the mechanisms that have enabled it to exert such influence on media and popular culture (Choudhury and Slaby, 2012b: 5–7). The goal of this article is to provide a small contribution to this vast critical endeavour, focusing particularly on the prominent modular hypothesis and its impact on the philosophical underpinnings of neuroscience and the interpretative pull it exerts on both the academic and the lay public.

Neuroscience: From Humanities to Animalities?

In the past few centuries, the development of natural sciences has brought about major changes in our everyday understanding of ourselves and the world. It is often contended that "the naïve self-love of man" has had to submit to three "major blows at the hand of science" (Freud, 1989: 353): First, Copernicus and Galileo demonstrated that the Earth is not the center of the universe; then, Darwin showed that animal species, including human beings, are the result of natural selection and not of divine design; and finally, Freud disclosed that much of our mental life is governed not by our free will, but by un- and sub-conscious processes. These three shifts have had a profound impact on cosmology, biology and psychology, but haven't completely changed or obliterated what Owen Flanagan (2002) refers to as "the humanistic image". Flanagan believes that the Western society has been dominated by two great worldviews: the humanistic and the scientific. The humanistic worldview consists of a set of beliefs about ourselves based on the assumption that we are spiritual beings with free will who are able to lead moral and meaningful lives. In contrast, the scientific worldview maintains that we have evolved according to the principles of natural selection and thus cannot circumvent the laws of nature.

The question is whether these two world-views are compatible. Flanagan suggests that they *can* coexist *if* we understand the humanistic image to reveal our spiritual nature and science as unlocking the secrets of the external world and our animal essence. He believes that this coexistence is not possible without the premise that we are *only partly* animal (Flanagan, 2002: xii). But the advances in evolutionary biology, cognitive science, and especially cognitive neuroscience cast shadow on this premise. The most avid proponents of the neuroscientific revolution are namely convinced that neuroscience-*cum*-evolutionary-psychology (referred to endearingly as "neuromania" and "darwinitis" by Tallis (2012)) is about to deliver the fourth, and possibly decisive, blow to the humanistic worldview. They feel that consciousness, cognition, and volition – the last surviving mysteries

(*cf.* Dennett, 1991) – can be fully explicated and accounted for in neurobiological terms. The more radical among the neuro-enthusiasts are even convinced that the concepts employed in the humanistic worldview (beliefs, emotions, free will, etc.) are mere illusions and have no reference in the (physical) world (e.g. Churchland, 1988; Wegner, 2002). This radical attitude finds a very vivid expression in Crick's words:

> *[T]he study of consciousness is a scientific problem. (...) There is no justification for the view that only philosophers can deal with it. Philosophers have had such a poor record over the last two thousand years that they would do better to show a certain modesty rather than the lofty superiority that they usually display. (...) I hope that more philosophers will learn enough about the brain to suggest ideas about how it works, but they also learn how to abandon their pet theories when the scientific evidence goes against them or they will only expose themselves to ridicule.* (Crick, 1994: 257–258)

Paraphrasing the famous words of Sellars, the view of the most fervent among advocates of the neuroscientific revolution might thus be summarised as follows: "[I]n the dimension of describing and explaining the world, [neuro]science is the measure of all things, of what is that it is and of what is not that it is not" (Sellars, 1963: 173).

But is neuroscience capable of living up to these bold claims? Bickle (1998, 2003) has argued that we should wait for scientific psychology and neuroscience to mature and only then examine to what extent it has managed to realise its goals. However, some of the advocates of the traditional humanistic world-view believe that the scientific image is suspect *in principle*, in that it leads to an impoverished and unnecessarily reductive image of the mind which cannot support the idea of a conscious and autonomous person capable of leading a moral and meaningful life:

> *Let us suppose we accept biologism in full: our minds are our brains; and our brains are evolved organs designed, as are all organs, by natural selection to maximize the replicative ability of the genes whose tool the brain is. What follows from this? (...) We may jettison the notion of freedom and, consequently, of personal responsibility. Worse still, to be identified with a piece of matter, and this, like all other pieces of matter, is subject to, and cannot escape from, the laws of material nature. (...) Our destiny, like that of pebbles and waterfalls, is to be predestined.* (Tallis, 2012: 51)

It seems that we are faced with two competing and incompatible approaches. But do we really have to abandon our intuitions of what it

means to be a human being if we treat mind as a natural phenomenon? In other words, does the acceptance of natural sciences necessarily force us to reduce the human worldview to the natural worldview? Is there any truth to Tallis' bitter remark: "If you want to understand people, look at their brains. The writing is on the wall and the script is pixels on a brain scan. Roll over, social sciences and humanities, allow yourselves to be incorporated into a vastly extended neuroscience and discover your true nature as animalities" (Tallis, 2012: 59)?

It is our contention that this is not the case. First of all, even if it were possible to find "a neural signature" for a given psychological or social phenomenon (e.g. free will or empathy), which is not very likely (see below), it is far from clear that this would result in hard determinism. Referring to the problem of free-will, Roskies contends that

> *[a] view of ourselves as biological mechanisms should not undermine our notion of ourselves as free and responsible agents. After all, some causal notion is needed for attributions of moral responsibility to make sense. The predictive power of our high-level psychological generalizations grounds our views of agency, so further evidence that we behave in a law-like fashion should not undermine our notions of freedom.* (Roskies, 2006: 421)

Secondly, and more importantly, it is questionable whether many of the far-reaching claims propagated under the banner of neuroscience are, in fact, scientifically and philosophically sound, i.e. whether they actually explain what they purport to explain. Quite the contrary, we believe that one of the important reasons for the current fascination with neuroscience is that it rests on a set of seductive, yet philosophically and scientifically suspect presuppositions, and that any serious discussion of the explanatory scope of neuroscience cannot afford to omit the analysis of these unreflected philosophical commitments (for a more in-depth account on the neuroscience and free will debate see Markič, 2009; Markič, 2011: chapter 6).

In what follows, we intend to focus on one of these presuppositions and delineate its major conceptual, methodological, and epistemological drawbacks. In this way, we hope to show that it is crucial not to take the neuroimage-based explanatory claims for granted – not all that glitters is gold, as the old saying goes –, but to subject them to rigorous and systematic investigation.

The Allure of Brain Imaging: This is Your Brain on X[2]

For some reason, dazzling and vibrant brain images seem to exert considerable authority on the public perception of scientific research findings. This is attested to by two recent studies on how the inclusion of neuroscientific information changes the appraisal of arguments and explanations.[3] In a study by Skolnick Weisberg et al. (2008), subjects were divided into three groups (lay adults, neuroscience students, and neuroscientists) and were given brief descriptions of psychological phenomena that are familiar from everyday experience. The descriptions were followed by one of four types of explanation: good/bad *neuroscientific* explanation or good/bad *non-neuroscientific* explanation. The study showed that subjects in all three groups judged good explanations as more satisfying than the bad ones, *except* when the explanations contained *irrelevant neuroscientific information* – in those cases, the two non-expert groups (lay adults and neuroscience students) tended to accept the bad explanation. In a similar study by McCabe and Castel (2008) three experiments are reported, demonstrating that the presence of brain images in an article (as compared to, say, bar graphs or no images) results in higher rating of scientific reasoning for arguments, i.e. "readers infer more scientific value for articles including brain images than those that do not, regardless of whether the article included reasoning errors or not" (351).

These results have sparked a host of criticism against neuroscience. The more fierce critics proclaimed neuroimages "a fast-acting solvent of critical faculties" (Crawford, 2010: 355), whose charm, novelty, and pictorial splendour tend "to overwhelm critical consideration" (Uttal, 2011: 21); to more reserved observers they were "epistemically compelling: [t]hey invite us to believe" (Roskies, 2010: 195). Note, however, that the fascination with neuroimagery is by no means limited to the lay public alone, but extends to (at least certain parts of) the neuroscientific community as well: the former is an enthusiastic consumer of "everything neuro", the latter its trustworthy supplier. Consider, for instance, two recent studies by Semir Zeki claiming to have found nothing less than the neurobiological underpinnings of *beauty* and *hate*. In the first study (Kawabata and Zeki, 2004) Zeki scanned subjects'

[2] *For reasons of space, we have decided to base our argument against the modular hypothesis exclusively on the "correlation problem", i.e. the issue of explicating the nature of a given mental phenomenon by means of brain images of neuronal activity in a specific brain region. Similar arguments would apply for clinical-pathological studies (e.g. attempts to draw conclusions on the nature of a specific mental phenomenon on account of its impairment following a lesion in a specific brain region) and stimulation techniques (e.g. attempts to draw conclusions on the nature of a specific mental phenomenon on account of its being actively brought about by direct electromagnetic stimulation of a specific brain region). For further details see Uttal 2001; 2012.*

[3] *But see Farah and Hook (2013) for a different reading.*

brains as they looked at pictures of people they hated (e.g. ex-lovers, work rivals etc.) and people about whom they felt neutrally; and in the second study (Zeki and Romaya, 2008) he scanned subjects' brains as they peered at pictures they had previously labelled as "ugly", "neutral", and "beautiful". By comparing brain activations elicited by the stimuli in experimental condition (hated faces and beautiful pictures, respectively) and those elicited by the stimuli in control condition (neutral faces and neutral/ugly pictures, respectively) Zeki concluded that he has identified neurobiological correlates of hatred and beauty. Let us, for argument's sake, assume that Zeki *had*, in fact, managed to pinpoint the exact neural correlates of hate and beauty (a bold and dubious statement, as we are about to see shortly) – the question we are immediately confronted with is *why* does this strike us as so compelling? *What* do the neuroimages *actually* tell us about hate and beauty? What is it about neural activation in a certain brain region that makes us believe that it may help us account for a host of seemingly complex and multifaceted experiential, mental, social etc. phenomena?

There is probably no uniform answer to this question, but one of the more prominent reasons for the obvious allure of neuroimages is the (implicit or explicit) acceptance of the *modular hypothesis*, i.e. the idea that the mind consists of a plethora of different cognitive functions (modules) and that these are somehow instantiated or realised in discrete brain regions (Tallis, 2012: 22–37; Uttal, 2001). So, why does the modular conception of the brain seem so appealing and why does it seem to hold such great promise for the explanation of mental and experiential phenomena? If we assume that specific brain regions are specialised for specific mental functions, then it seems that mental properties of a certain experience could be analytically explained (away?) by the (say, causal) properties of specific brain regions. If, for instance, it turns out that, as suggested by Zeki, the orbito-frontal cortex is involved in the processing of beautiful pictures, then it would seem plausible that the experience of beauty that is commonly assumed to be the basic of aesthetics might be accounted for in terms of the activation of this particular region. In other words, the modular conception of the brain rests on the idea that properties of a given mental phenomenon are nothing but the sum total of properties of brain regions that have been shown to accompany this experience. Thus, in a recent study, Meeks and Jeste (2009) propose a "speculative model of the neurobiology of wisdom [!]", observing that

> *the prefrontal cortex figures prominently in several wisdom subcomponents (e.g. emotional regulation, decision making, value relativism) primarily via top-down regulation of the limbic and striatal regions. The lateral prefrontal cortex facilitates calculated, reason-based decision*

> *making, whereas the medial prefrontal cortex is implicated in emotional valence and prosocial attitudes, behaviours. Reward neurocircuitry (ventral striatum, nucleus accumbens) also appears important for promoting prosocial attitudes/behaviours.* (Meeks and Jeste, 2009: 355)

The rationale is pretty straightforward: "Wisdom" is first analytically (and rather provisionally) divided into its basic subcomponents, and these are then correlated with, and explained by, the activity in the corresponding brain regions. The overall circuitry – the sum total of all interconnections between relevant brain regions instantiating individual subcomponents – is then believed to provide us with a (speculative) neurobiological account of wisdom.

Yet the modularity thesis is problematic for several reasons. The first reason is that it doesn't seem to correspond well with how the brain actually functions:

> *Studies that suggest 'a brain spot for X' are typically misleading because mental functions are rarely localized to one place of the brain. There is a Babel of crosstalk among numerous regions as they are strung together in specialized neural circuits that work in parallel to process thoughts and feelings.* (Satel and Lilienfeld, 2013: 15–16)

The brain is "a dynamic, functionally integrated, and highly interdependent system of complex synaptic-neural networks that interact in non-linear ways"; in such a system there is no isolated neural activity, as each activity, even if it might *seem to be* distinct from other happenings in the brain, is actually a part of an integrated mesh of broader circuitry (Cunningham, 2011: 228). Cunningham[4] (2011) provides a comprehensive list of neuroscientific findings substantiating this claim. First, the seemingly discrete brain regions actually merge seamlessly with other regions (e.g. there is a strong overlap of various sensory and motor regions; *cf.* Uttal, 2001: 159). Second, areas that seem to be functionally demarcated activate broadly distributed brain regions (e.g. speech areas; *cf.* Uttal, 2001: 154–155). Third, certain brain regions (e.g. cerebellum) that were once considered to be involved in only one function (e.g. motor coordination) are now known to perform several functions (e.g. language processing, problem solving, and memory tasks; *cf.* Uttal, 2001: 158). Fourth, the same neural network can perform different functions and different networks can perform similar functions (the so-called *multiplexing*). Fifth, the phenomenon of recovery of a function

[4] *It should be noted that as much as we agree with Cunningham's criticism of some of the unfounded presuppositions of neuroscience, we find his alternative dualist proposal ("the mediatory brain" hypothesis) unconvincing and metaphysically moot.*

(after a stroke, etc.) indicates that dynamical alterations in the localization are possible and that the brain can plastically reorganise itself (*cf.* Schwartz and Begley, 2002).

The second problem bedevilling the modular thesis is related to technical and methodological issues. First of all, it should be noted that fMRI *doesn't* measure the brain activity directly, but only *indirectly* by measuring *increases in the blood flow*. The underlying principle is that increased brain activity is correlated with increased metabolism and consequentially with increased oxygen consumption and blood flow. So, the key to registering brain activity is the detection of "the increases in blood flow needed to deliver oxygen to busy neurons" (Tallis, 2012: 76). What might seem like a triviality, however, can have tremendous impact on the interpretative results. Tallis explains:

> *Given that neuronal activity lasts milliseconds, while detected changes in blood flow lag by 2–10 seconds, it is possible that the blood flow changes may be providing oxygen to more than one set of neuronal discharges. What is more, many millions of neurons have to be activated for a change in blood flow to be detected. Small groups of neurons whose activity elicits little change in blood, or a modest network of neurons linking large regions, or neurons acting more efficiently than others, may be of great importance but would be under-represented in the scan or not represented at all.* (Tallis, 2012: 76)

Secondly, and more importantly, neuroimages are the result of *subtraction* and *statistical averaging*: A baseline measurement of brain activity in the control condition (e.g. looking at a neutral picture) is subtracted from a measurement of brain activity in the experimental condition (e.g. looking at a beautiful picture) and the obtained differential image (the assumed neurobiological substratum of the sense of beauty) is then statistically analysed to filter out the background noise and average the results across all participants in the study (Crawford, 2010: 360). The resemblance between brain scans and pictures is therefore deceiving:

> *Photos capture images in real time and space. Functional images are constructed from information derived from the magnetic properties of blood flowing in the brain. (...) [A]t their most accurate, they simply represent local activation based on statistical differences in [the measured oxygenation levels].* (Satel and Lilienfeld, 2013: 7)

In other words, the dazzling fMRI neuroimages *don't* correspond to the *actual activity* in well-defined brain regions, but are to a considerable

degree the result of measuring and statistical techniques that are employed in brain imaging procedures. These techniques obscure the fact that much more of the brain is active in both conditions and gives the false impression that the activation is restricted to a neatly circumscribed brain region.

The third problem of the modular hypothesis pertains to its explanatory vacuity and circularity. To get a better grasp of the issue, let us try to reconstruct how a neuroscientist might come to a conclusion that a certain brain region is associated with the such-and-such cognitive function. It is important to note that her realization *wasn't* derived from a careful investigation of individual brain regions *per se*, because there are no *intrinsic* neural properties that would, in themselves, explain why it is *precisely this* region that plays a key role in *precisely this* function/state. On the contrary, the neuroscientist has come to her conclusion *correlatively*, i.e. by drawing parallels between changes occurring in mental functions/states and the corresponding changes in neuron activity. When she has collected a sufficient amount of such correspondences, she is able to draw a conclusion that this region is *somehow* associated with a such-and-such mental state/function. It is true that, once the correspondence has been established, a scientist might *retrospectively* attempt to find (tentative) reasons about how and why a certain area may contribute to a given conscious phenomenon, but such attempts are possible only *post festum*, as the brain tissue itself *is silent*: it is futile to try and guess what a function of a certain brain region might be if there are no psychophysical indications of what that region is supposed to be doing:

> *Even if it is an attempt to capture what is objectively happening inside the brain, an fMRI or PET scan lacks any pertinence for the study of consciousness [or the mind] unless it is correlated to the subject's first-person experience. Indeed, the only reason brain states or functional states assume the relevant importance they do is through their putative correlation with mental sates identified on other, experiential grounds.* (Gallagher and Zahavi, 2008: 16)

Now, what is it we would learn if it turned out that a given mental state *M* is, in fact, always accompanied by an activation of a certain neural network *N*? Given that the function of *N* has *previously* been determined on the basis of (behavioural or verbal) accounts of mental and experiential states accompanying *N*, it is unclear what – *except for the stamp of scientific authenticity* – the mentioning of the neurobiological level would actually contribute to the understanding of *M*:

> *Our only knowledge of the functional architecture of the brain stems from the collection of reports and behavioral observations with which*

> *the measured brain activities are correlated. Looking at just the brain reveals nothing interesting to cognitive scientist. […] The brain areas that have come to be known as "sensory areas" have only been recognised as such based on a collection of reports about sensations.* (Overgaard, 2004: 370)

Satel and Lilienfeld coined the term "neuroredundancy" to denote "things we already knew without brain scanning". This neural version of *nihil novi sub sole* is nicely exemplified by Paul Zak's enthusiastic claim that a brain scan lets us "embrace words like 'morality' or 'love' or 'compassion' in a non squishy way. These are *real things*" (Satel and Lilienfeld, 2013: 21–22; our emphasis). Similarly, neuroscientist Andrew Newberg is convinced that by having elucidated the neurobiological underpinnings of religious experience, he has managed to show that these experiences are "real" (d'Aquili and Newberg, 1999; Newberg and d'Aquili, 2002). But what exactly does that mean? What "non squishy reality" is conferred upon "morality", "love", "compassion", and "mystical experience" by a surplus of vibrant brain imagery?

Neuroredundancy can thus be seen as an offshoot of a broader and more serious phenomenon termed by Racine as "neurorealism". Neurorealism refers to "the misbegotten propensity to regard brain images as inherently more 'real' or valid than other types of behavioral data" (Satel and Lilienfeld, 2013: 21), and might have important implications on how we evaluate and interpret experimental results. For one thing, it legitimises the dubious process of the so-called "reverse inference" (Poldrack, 2006) where we try to reason backward from neural activity to subjective experience:

> *The difficulty with reverse inference is that specific brain structures rarely perform single tasks, so one-to-one mapping between a given region and a particular mental state is nearly impossible. In short, we can't glibly reason backward from brain activations to mental functions.* (Satel and Lilienfeld, 2013: 13)

Moreover, neurorealism might forcibly and unfoundedly prioritise neural accounts over all other accounts. Let us look at two examples. First, a study by Eisenberger et al. (2003) showed that there is an extensive overlap between brain regions implicated in physical and social pain. The authors concluded that the two types of pain are basically identical, claiming that humans are social animals and that the need for social cohesion demands social exclusion to be painful, which is achieved by employing brain circuitry already implicated in physical pain (Tallis, 2012: 79–80). In a similar manner, Gelbard-Sagiv et al. (2008) found that the same neuronal network

was activated when subjects remembered a scene from *The Simpsons* as when they actually saw it, and therefore inferred that memory involves the reactivation of neuronal pathways that are activated in vision (Gelbard-Sagiv et al., 2008: 129–130). The problem is that the phenomena that are conflated on a *neuroscientific* level seem to be very different on a *phenomenological* and *psychological* level. Should we therefore conclude that our experiential and psychological conceptions of the nature of physical and social pain or memory and perception are necessarily misguided? Or is, as Tallis argues convincingly, "the failure to demonstrate fundamental differences between what you feel when you stub your toe and your feelings when you are blackballed by a club from which you are seeking membership *a measure of the limitations of fMRI scanning* and, indeed, other modes of brain scanning" (Tallis, 2012: 80; our emphasis)? What reasons are there, aside from the just-so evolutionary story provided by Eisenberger et al., to negate other levels of description and explanation in favour of the purportedly more scientific description and explanation provided by neuroscience?

The fourth problem with the modular hypothesis, and one that is closely related to the issue of neurorealism, is that it often conflates three essentially different relations: correlation, causation, and identity. An important reason why finding a neural correlate *N* of a mental phenomena *M* seems so compelling is that it fosters the (false) impression of revealing *the true cause* or *the "true (physical) nature"* of *M*. In other words, it is often uncritically assumed that since mental and experiential events are accompanied by neural events, the former can be *accounted for* or *reduced to* the latter. In Rockwell's words:

> *For those who do neuroscience, it is highly effective to assume that brain events are "the" cause of mental events. There is overwhelming empirical evidence that whenever a mental event occurs, something happens in the brain. Conversely, when something happens to the brain, it frequently has an effect on the mental events of the person who possesses that brain. The omnipresence of these reciprocal causal connections has prompted the natural assumption that the mind is the brain.* (Rockwell, 2007: 54)

In a philosophically naive sleight of hand, "*M* is accompanied by *N*" becomes "*M* is caused by *N*" or "*M* is reducible to *N*", despite the fact that we have currently absolutely no idea of how and why it would be possible for *N* to either cause *M* or even be identical with *M*:

> *Even if we could find precise modular locations in the brain associated with well-defined psychological constructs, we still would not have solved*

> *the problem of how brain activity becomes mental activity.* (Uttal, 2001: 70, 126)

This brings us to the so-called "hard problem of consciousness", the problem of conscious experience or phenomenal consciousness (also referred to as qualia in philosophy):

> *The really hard problem of consciousness is the problem of [conscious] experience. When we think and perceive, there is a whir of information-processing, but there is also a subjective aspect. As Nagel (1974) has put it, there is something it is like to be a conscious organism. This subjective aspect is experience. When we see, for example, we experience visual sensations: the felt quality of redness, the experience of dark and light, the quality of depth in a visual field. (...) It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does.* (Chalmers, 1995: 201)

M (at least its experiential aspects) and *N* are separated by the notorious *explanatory gap* (Levine, 2002): No matter how exact or thorough our knowledge of *N* that accompanies *M*, it will fail to provide a satisfactory explanation as to why *M* is the way it is or why *M* is accompanied by *N*. In other words, the *correlates* between neurobiological and experiential states, in themselves, tell us nothing about *how* these states are *mutually interrelated* (causality, identity, duality, etc.), so an attempt to account for the nature of this relationship is not a scientific, but a *metaphysical* question (see also Strle, 2013; Kordeš, 2013).

From Science to Scientism: Phrenology and Neurotheology

Recently, several authors (e.g. Dobbs, 2005; Kuran, 2011; Tallis, 2012; Uttal, 2001) have proposed that modular approaches in neuroscience seem to be dangerously drifting towards phrenology, a controversial (pseudo) scientific discipline from the 19th century. Although we partly agree with Satel and Liliengfeld, who point out that such analogies tend to be somewhat exaggerated (2013: 3), we would like to shortly delineate some of the reasons why they might nonetheless contain a grain of truth. We would like to demonstrate this by comparing phrenology with one of the more recent and arguably the most notorious neuro-discipline, neurotheology.

Neurotheology emerged in the 1970's and 1980's, but it wasn't until the end of the 1990's that it started to gain more attention *and* notoriety, not least due to

fashionable catchphrases, such as the "God Machine" (Horgan, 2004: 91–105) and the "God Part of the Brain" (Alper, 2006) that have become associated with it. The main goal of neurotheology is to account for the phenomenon of religion in neuroscientific terms, and it sets out to do this by studying primarily *religious experience*: unlike other aspects of religion (e.g. ritual, belief, etc.), its experiential aspects fall under the purview of experimental neuroscience (i.e. they are amenable to brain imaging studies), so they seems to be the perfect starting point to account for how religious phenomenology arises from neuropsychology (d'Aquili and Newberg, 1999). The field of neurotheology is extremely diverse, and there is very little, if any, consensus among individual authors on which brain regions are implicated in religious experience. In the past few decades, numerous neurobiological models of religious experience have been put forward: from (a) *unimodular* – right-hemisphere hypothesis (Ornstein, 1972) and temporal-lobe hypothesis (Ramachandran and Blakeslee, 1998; Persinger, 1983; Persinger and Healey, 2002) – through (b) *bimodular* – temoral-parietal-lobe hypothesis (d'Aquili and Newberg, 1999; Newberg and d'Aquili, 2001) and temporal-frontal-lobe hypothesis (McNamara, 2009) – to (c) *multimodular* or *systemic* models (Austin, 1999, 2006; Beauregard and O'Leary, 2007). All these models have been criticised from empirical, methodological, conceptual, and philosophical perspectives (for a more in-depth account see Vörös, 2010; Vörös, 2012; Vörös, 2013: chapter 2).

Phrenology, on the other hand, was developed at the turn of the 19th century by Franz Joseph Gall (1758–1828) and Johann Spurzheim (1776–1832) and was based on three fundamental postulates: (a) the brain is the *organ of the mind*; (b) the functions of the brain are *modular*, i.e. different brain regions are responsible for different mental functions; and (c) precise *measurements of the skull* can reveal the extent to which a given brain module, and consequently the corresponding mental faculty, is developed (Tallis, 2012: 33). For instance, Franz Joseph Gall believed that mental functions are localised in discrete parts of the brain, which he referred to as "organs", and that each of these organs (modules) was a substrate of a particular mental faculty. He was also convinced that the functional strength of a given cerebral organ was determined by its volume, and that the latter, in turn, determines the correlative size of the bulges and bumps in the region of the skull adjacent to a given cerebral organ. Gall thus maintained that, by observing, palpitating and measuring the skull, it would be possible to construct a map of brain organs that instantiate different psychophysical features (Kuran, 2011; Simpson, 2005).

Why was phrenology eventually discounted as *pseudo*-science? Note that, contrary to first appearances, the main reason was not so much its methodology, as the fact that it had unwarrantedly transcended its basic theoretical framework and attempted to provide answers (explanations) to

questions that were out of its bounds. Even more importantly, all this was done *in the name of science*: "Because phrenology was seen as based on scientific 'facts', advocates used this authority to make claims about issues far removed from phrenology" (Norman and Jeeves, 2010: 236). The modern reader might chuckle upon learning what kind of "empirical data" Gall's *Schädellehre* (skull reading) used to substantiate its claims (bumps in the skull, etc.), but we should pause to wonder whether the situation will be truly that different when people living 100–200 years from now are confronted with the findings and methods of contemporary neuro-disciplines.

At its inception, phrenology was a *perfectly legitimate* and even *highly original* research programme. Its main problems weren't its theoretical assumptions *as such* – these were formulated in the form of (interesting) *hypotheses*, which the later empirical research might have corroborated or (as it actually happened) undermined –, but the fact that phrenology, operating under the pretext of "doing strict science", used these (untested!) assumptions to account for highly complex psychological, sociological, cultural, and religious aspects of human life (from temperament and personality to morality and religion). What made phrenology *non*-scientific or *scientistic*[5], and eventually contributed to its progressive decline, was therefore not so much the result of the non-"empirical" data that it used to substantiate its claims with – in its early stages there was no way of telling whether these might prove reliable or not – as the fact that it tried to deduce *scientific (empirically corroborated) answers* to the above-mentioned questions on the grounds of *uncorroborated (theoretical) hypotheses*. Phrenology became ideology at the very moment it tried to create an impression that its assumptions weren't *speculative*, but *scientific*. And what connects phrenology with modern neurotheology, is precisely the (uncorroborated!) *assumption of modularity*:

> *[A]lthough the 'bumps on the skull' idea is no longer with us, the idea that mental components exist and that they can be assigned to specific locations of the brain very much is. Indeed, the central problem facing cognitive neuroscience is how to deal with the unproven assumption that mental processes are accessible, separable, and localizable as are the material aspects of the brain.* (Uttal, 2001: 108–109)

The *data collecting techniques* might have changed – skull measurements have been replaced by brain scans –, but the *basic background assumption*

[5] *Following Crawford, "scientism" can be understood as "the overextension of some mode of scientific explanation, or model, to domains in which it has little predictive or explanatory power" (Crawford, 2010: 356).*

remains the same, and we have seen this assumption to be problematic for several reasons. Hence, one can but agree with Bradford who, in his critical assessments of neurotheology, sees neuroimaging techniques as "a boon" for neuroscience, as they confer its claims with "a halo of certainty" (Bradford, 2012: 111). Just as observations of the skull, measurements of bulges etc. were used in the past to provide phrenology with an aura of "scientific credibility", so too are the modern imaging techniques, such as fMRI, PET or SPECT, all too often (mis)used and (mis)portrayed by certain members of the neuroscientific community to embellish their findings with a "stamp of scientific authenticity". Vivid neuroimages are not "facts", but "data" in need of interpretation. Empirical findings *in themselves* are *silent*; and whoever disdains "philosophising" and demands of empirical findings to speak *for themselves*, does little more than obscure the metaphysical presuppositions on which one's claims are based.

The broad and colourful spectrum of background metaphysical presuppositions discloses itself once we consider how phrenological and neurotheological authors deal with questions about the nature and reliability of religious experience. As pointed out by Norman and Jeeves, the theoretical framework of phrenology was compatible with a whole spectrum of different metaphysical positions. Some phrenologists were convinced that phrenology will "replace" religion, some that it will "purify" religion, and others that it will "harmonize" religion with science. Some maintained that the "soul" was "using the brain", others that it was merely "its manifestation"; "revelation" was believed to be "superior" to scientific truths by some, and "inferior" by others (Norman and Jeeves, 2010: 239). An almost identical situation – "the same diversity of opinions" – is present in modern neurotheology: some authors maintain a "materialist position", some are "non-materialists", and others are "noncommittal" (Norman and Jeeves, 2010: 243). What all these examples have in common, however, is the fact that the positioning of individual authors *isn't* based on empirical findings, but on their *implicit/background (metaphysical) presuppositions*, which normally remain *unreflected*. That is why what might have ended up as a potentially interesting scientific discipline turned not only into *bad science*, but also into *bad philosophy (scientism)*.

Conclusion: Towards Critical Neuroscience

It has been argued throughout this paper that *certain currents* within the neuroscientific community, in their disregard of fundamental conceptual, methodological, and epistemological presuppositions of their discipline, rush head-on into the quagmire of old, and mostly superseded, metaphysical befuddlements. This epistemic myopia is especially troubling, as it threatens

to repeat the same mistakes that had brought about the spectacular, yet informative downfall of phrenology in the 19th century. The numerous *ad hoc* "neurologisations" of psychological and social life might increase the popularity of neuroscience in the short run, but at the high price of devaluing its scientific credibility in the long run.

Yet despite first appearances, the main point of this article is not to spark a backlash of fury from the embittered humanities, but to pave the way towards a more integrated conception of neuroscience, one "that situates the brain and cognition in the body, the social milieu, and the political world" (Choudhury and Slaby, 2012b: 3). To initiate this process, however, it is crucial that the major theoretical obstacles are brought to light and that neuroscience itself initiates a (self-)transformative process. Methodological reductionism and modularism might be an indispensable *experimental tool* for neuroscience, but they become dubious if their explanatory powers are over-extended. Humans are embodied and cultural beings embedded in complex cultural and historical frameworks, that is why sciences of the mind need to take into account not only happenings in the brain, but also in other somatic, cultural, political, etc. systems. The reckless quest for precise neurobiological correlates of mentality and consciousness, i.e. the attempt to capture the neurobiological "photograph of the soul", is therefore methodologically, conceptually, and epistemologically suspect. The explanatory framework does not consist solely of a one-way traffic from neuroscience to higher level sciences, but also of an *opposite* movement: Neurobiological research might shed light on certain aspects of mental and social phenomena, which, in turn, help us understand the nature and scope of neurobiological research (see also Markič, 2013). This aligns nicely with Walter's characterisation of neurophilosophy

> *as a discipline that moves in on the mind-brain problem from two opposite directions. Either we begin on the empirical side and happen upon philosophical questions, or we set out with philosophical puzzles and need empirical findings to solve them. (...) It is best understood as a bridge discipline between subjective, experiential, philosophical theorising, and empirical research.* (Walter, 2001: 25)

Similar ideas have been proposed under the heading of neurophenomenology (Varela, 1996) and embodied/enactive cognitive science (Varela et al. 1991; Thompson 2007, Ward and Stapleton, 2012). The goal is not to undermine the neuroscientific endeavour, but to externally contextualise and internally solidify it. This, however, entails a critical reassessment of some of the bold and far-reaching claims being made in the name of neuroscience and putting the burgeoning field of neuro-disciplines into the

philosophical, social, historical, and political perspective. In other words, it is important to establish a fruitful platform for criticism that is "constructive and engaged with neuroscientific research" (Choudhury and Slaby, 2012b: 3), and thus enables a dynamic two-way dialogue between neuroscience and other areas of research.

BIBLIOGRAPHY

Alper, Matthew (2008): The "God" Part of the Brain: A Scientific Interpretation of Human Spirituality and God. Naperville: Sourcebooks.

Austin, James H. (1999): Zen and the Brain. Cambridge, MA, London: MIT Press.

Austin, James H. (2006): Zen-Brain Reflections. Cambridge, MA, London: MIT Press.

Beauregard, Mario and O'Leary, Denyse (2007): The Spiritual Brain. A Neuroscientists Case for the Existence of the Soul. New York: Harper Collins.

Bennett, Maxwell R. and Peter M. S. Hacker (2003): Philosophical Foundations of Neuroscience. Molden, MA, Oxford: Blackwell Publishing.

Bickle, John (1998): Psychoneural Reduction: The New Wave. Cambridge, MA, London: MIT Press.

Bickle, John (2003): Philosophy and Neuroscience: A Ruthlessly Reductive Account. Dordrecht: Kluwer Academic Publishers.

Bradford, David T. (2012): A Critique of 'Neurotheology' and an Examination of Spatial Perception in Mystical Experience. Acta Neuropsychologica 10 (1): 109–123.

Chalmers, David J. (1995): Facing Up the Problem of Consciousness. Journal of Consciousness Studies 2 (3): 200–219.

Choudhury, Suparna and Jan Slaby (eds.) (2012a): Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience. Chichester: Blackwell Publishing Ltd.

Choudhury, Suparna and Jan Slaby (2012b): Introduction: Critical Neuroscience – Between Lifeworld and Laboratory. In Suparna Choudhury and Jan Slaby (eds.), Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience, 1–26. Chichester: Blackwell Publishing Ltd.

Churchland, Paul M. (1988): Matter and Consciousness. Cambridge, MA: MIT Press.

Crick, Francis (1994): The Astonishing Hypothesis: The Scientific Search for the Soul. New York: Scribner.

Crawford, Matthew, B. (2008/2010): The limits of neuro-talk. In Giordano, James J. and Bert Gordijn (eds.), Scientific and Philosophical Perspectives in Neuroethics, 355–369. Cambridge: Cambridge University Press.

Cunningham, Paul F. (2011): Are Religious Experiences Really Located Within the Brain? The Promise, Challenges and Prospects of Neurotheology. Journal of Mind and Behavior 32 (3): 223–249.

d'Aquili, Eeugene and Andrew B. Newberg (1999): The Mystical Mind: Probing the Biology of Religious Experience. Minneapolis: Fortress Press.

Dennett, Daniel C. (1991): Consciousness Explained. Boston: Little, Brown and Company.

Dobbs, David (2005): Fact or Phrenology?. Scientific American Mind 1 (1): 24–31.
Eisenberger, Naomi I., Matthew D. Liebermann and Kipling D. Williams (2003): Does Rejection Hurt? An fMRI Study of Social Exclusion. Science 302 (5643): 290–293.
Farah, Martha J. and Cayce J. Hook (2013): The Seductive Allure of 'Seductive Allure'. Perspectives on Psychological Science 8 (1): 88–90.
Flanagan, Owen (2002): The Problem of the Soul: Two Visions of Mind and How to Reconcile Them. New York: Basic Books.
Freud, Sigmund (1989): Introductory Lectures on Psychoanalysis. New York: Norton.
Gallagher, Shaun and Dan Zahavi (2008): The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science. London, New York: Routledge.
Gebard-Sagiv, H., Mukamel R., Harel M., Malach R. and Fried I. (2008): Internally Generated Reactivation of Single Neurons in Human Hippocampus During Free Recall. Science 322 (5898): 96–101.
Horgan, John (2004): Rational Mysticism: Spirituality Meets Science in the Search of Enlightenment. Boston, New York: Mariner Books.
Illes, Judy, Mary Anne Moser, Jennifer B. McCormick, Eric Racine, Sandra Blakeslee, Arthur Caplan, Erika Check Hayden, Jay Ingram, Tiffany Lohwater, Peter McKnight, Christie Nicholson, Anthony Phillips, Kevin D. Sauvé, Elaine Snell and Samuel Weiss (2010): Neurotalk: Improving the Communication of Neuroscience Research. Nature Reviews Neuroscience 11(1): 61–69.
Kawabata, Hideaki and Semir Zeki (2004): Neural Correlates of Beauty. Journal of Neurophysiology 91 (4): 1699–1705.
Kordeš, Urban (2013): Problems and opportunities of first-person research. Interdisciplinary Description of Complex Systems, 11(4): 363–375.
Kuran, Manuel (2011): Nevroteologija med frenonologijo in nevromitologijo. Časopis za kritiko znanosti 29 (246): 36–50.
Levine, Joseph (1983/2002): Materialism and Qualia: Explanatory Gap. In David John Chalmers (ed.), Philosophy of Mind: Classical and Contemporary Readings, 354–361. Oxford University Press, New York, Oxford.
Lynch, Zack (2009): The Neuro Revolution: How Brain Science is Changing the World. New York: St. Martin's Press.
Markič, Olga (2009): Neuroscience and the Image of the Mind. In Eva Žerovnik, Olga Markič and Andrej Ule (eds.), Philosophical Insights About Modern Science, 135–144. New York: Nova Science Publishers, Inc.
Markič, Olga (2011): Kognitivna znanost: filozofska vprašanja. Maribor: Aristej.
Markič, Olga (2013): The philosophical framework for understanding neuroscientific research. Interdisciplinary Description of Complex Systems, 11(4): 351–362.
McCabe, David P. and Alan D. Castel (2008): Seeing is Believing: The Effect of Brain Images on Judgement of Scientific Reasoning. Cognition 197: 343–352.
McNamara, Patrick (2009): The Neuroscience of Religious Experience. New York: Cambridge University Press.

Meeks, Thomas and Dilip Jeste (2009): Neurobiology of Wisdom: A Literature Overview. Archives of General Psychiatry 66 (4): 355–365.
Newberg, Andrew B. (2010): Principles of Neurotheology. Farnham, Burlington: Ashgate.
Newberg, Andrew B. and Eugene d'Aquili (2002): Why God Won't Go Away. New York: Ballantine Books.
Norman, Wayne D. in Malcolm A. Jeeves (2010): Neurotheology: Avoiding a Reinvented Phrenology. Perspectives on Science and Christian Faith 62 (4): 235–251.
Ornstein, Robert E. (1972): The Psychology of Consciousness. Harmondsworth etc.: Penguin Books.
Overgaard, Morten (2004): On Naturalising of Phenomenology. Phenomenology and the Cognitive Science 3: 365–379.
Persinger, Michael A. (1987): Neuropsychological Bases of God Beliefs. Westport: Praeger.
Persinger, Michael A. and Faye Healey (2002): Experimental Facilitation of the Sensed Presence: Possible Intercalation between the Hemispheres Induced by Complex Magnetic Fields. The Journal of Nervous and Mental Disease 190 (8): 533–541.
Persinger, Michael A. (1983): Religious and Mystical Experiences as Artefacts of Temporal Lobe Function: A General Hypothesis. Perceptual and Motor Skills 57 (3): 1255–62.

Poldrack, Russell (2006): Can Cognitive Processes Be Inferred from Neuroimaging Data? Trends in Cognitive Science 10 (2): 59–63.
Ramachandran, Vilayanur S. and Sandra Blakeslee (1998): Phantoms in the Brain. New York, London, Toronto, Sydney: Harper Perennial.
Rockwell, Teed (2007): Neither Brain nor Ghost. Cambridge, MA: MIT Press.
Roskies, Adina (2006): Neuroscientific challenges to free will and responsibility. Trends in Cognitive Sciences 10 (9), 419–423.
Roskies, Adina (2010): Are Neuroimages like Photographs of the Brain? Philosophy of Science 74: 860–872.
Satel, Sally and Scott O. Lilienfeld (2013): Brainwashed: The Seductive Appeal of Mindless Neuroscience. New York: Basic Books.
Schwartz, Jeffrey M. and Sharon Begley (2002): The Mind & the Brain: Neuroplasticity and the Power of Mental Force. New York etc.: Harper Perennial.
Sellars, Wilfrid (1963): Science, Perception and Reality. London: Routledge.
Simpson, D. (2005): Phrenology and the Neurosciences: Contributions of F. J. Gall and J. G. Spurzheim. Autralian and New Zealand Journal of Surgery 75: 475–482.
Skolnick Weisberg, Deena, Frank C. Keil, Joshua Goodstein, Elizabeth Rawson and Jeremy R. Gray (2008): The Seductive Allure of Neuroscience Explanations. Journal of Cognitive Neuroscience 20 (3): 470–477.
Strle, Tomaž (2013): Why should we study experience more systematically: neurophenomenology and modern cognitive science. Interdisciplinary Description of Complex Systems, 11(4): 376–390.

Tallis, Raymond (2012): Aping Mankind: Neuromania, Darwinitis and the Misrepresentation of Humanity. Durham: Acumen.
Thompson, Evan (2007): Mind in Life: Biology, Phenomenology, and the Sciences of the Mind. Harvard University Press, Cambridge, MA.
Uttal, William R. (2001): The New Phrenology. Cambridge, MA: MIT Press.
Uttal, William R. (2011): Mind and Brain: A Critical Appraisal of Cognitive Neuroscience. Cambridge, MA: MIT Press.
Varela, Francisco J. (1996): Neurophenomenology: A Methodological Remedy for the Hard Problem. Journal of Consciousness Studies 3 (4): 330–349.
Varela, Francisco J., Eva Thompson and Rosch Eleanor (1991): The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Cambridge, MA.
Vörös, Sebastjan (2010): Blišč in beda nevroteologije. Analiza 9 (1–2): 71–92.
Vörös, Sebastjan (2012): Mit o božanskem senčnem režnju: učna ura v nevrofilozofiji. Anthropos 44 (3–4): 79–100.
Vörös, Sebastjan (2013): Podobe neupodobljivega: (nevro)znanost, fenomenologija, mistika. Ljubljana: KUD Logos.
Walter, Henrik (2001): Neurophilosophy of Free Will: From Libertarian Illusions to a Concept of Natural Autonomy. Cambridge, Ma., London: MIT Press.
Ward, Dave and Mag Stapleton (2012): Es are good: Cognition as enacted, embodied, embedded, affective and extended. In Fabio Paglieri (ed.), Consciousness in Interaction: The role of the natural and social context in shaping consciousness, 89–104. Amsterdam: John Benjamins Publishing Company.
Wegner, Daniel (2002): The Illusion of Conscious Will. Cambridge, Ma., London: MIT Press.
Zeki, Semir and J. P. Romaya (2008): Neural Correlates of Hate. PLoS One 3 (10): 35–56.