Intentional concepts in cognitive neuroscience

Samuli Pöyhönen[*], University of Helsinki
samuli.poyhonen@helsinki.fi

Penultimate draft, 3.9.2012
Forthcoming in *Philosophical Explorations*

Abstract

The paper develops an account of the use of intentional predicates in cognitive neuroscience explanations. As pointed out by Maxwell Bennett and Peter Hacker, intentional language abounds in neuroscience theories. According to Bennett and Hacker, the subpersonal use of intentional predicates results in conceptual confusion. I argue against this overly strong conclusion by evaluating the contested language use in light of its explanatory function. By employing conceptual resources from the contemporary philosophy of science, I show that although the use of intentional predicates in mechanistic explanations sometimes leads to explanatorily inert claims, intentional predicates can also successfully feature in mechanistic explanations as tools for the functional analysis of the explanandum phenomenon. The approach developed in the paper also suggests a more general strategy for answering the question of what kind of language can be employed in mechanistic explanations.

Keywords: cognitive neuroscience, explanation, mechanism, intentionality

[*] An early draft of the article was presented at the ESPP 2010 meeting of the European Society for Philosophy and Psychology (Bochum, August 2010). I thank the audience of this event as well as the philosophy of science group and members of the 'A science of the soul' project at the University of Helsinki for their helpful comments. In particular, I am grateful to Olle Blomberg, Andy Clark, Jaakko Kuorikoski, Lisa Muszynski, Uskali Mäki, Petri Ylikoski and the two anonymous referees for this journal for reading and commenting on earlier versions of this paper.

1 Introduction

In this paper I develop an account of the use of intentional predicates in mechanistic neuroscience explanations. There has been extensive discussion on mechanisms in the recent philosophy of science (Machamer et al. 2000; Glennan 2005; Bechtel & Abrahamsen 2005). However, the discussion has mainly focused on ontological questions about the nature of mechanisms and their parts, and epistemic considerations regarding the discovery of mechanisms and their explanatory role. The question of how mechanisms ought to be (linguistically) described has received considerably less attention. However, this question is at the heart of the debate around Maxwell Bennett and Peter Hacker's (henceforth "B&H") well-known critique of cognitive neuroscience (Bennett & Hacker 2003, 2007, 2008). B&H lay out a wealth of examples from psychological and neuroscience research showing that in the mind sciences it is common to describe the functioning of the brain and its parts with intentional predicates. The famous claim of the authors is that in describing neural and cognitive structures in intentional terms – as believing something, making inferences, or forming hypotheses, to mention just a few examples – neuroscientists make a conceptual mistake that leads to nonsense and nullifies the explanatory power of their theories.

The reactions to B&H's work have been strongly polarized. Several authors have questioned the plausibility of the whole philosophical framework from which B&H present their critique of cognitive neuroscience (Churchland 2005; Dainton 2007; Dennett 2007; Machamer & Sytsma 2009). And as has been pointed out by their critics, B&H's approach indeed has implausibly strong implications: many of the research programs that they brand as nonsensical are among the most successful theories in contemporary scientific psychology.[1] Such strong implications might indeed serve as a reductio against their position. On the other hand, several favorable reviews both in philosophy and in the human sciences have considered B&H's writings an important contribution to the research on the conceptual foundations of cognitive neuroscience (Kenny 2008; Patterson 2003; Slaney & Maraun 2005; Schaal 2005). I interpret these positive reactions as a sign of B&H's argument appealing to some plausible intuitions, and therefore being worthy of more careful study.

However, it is not my aim to construct a careful interpretation of B&H's position in this paper. There are two reasons for why I take their critique as the starting point for my account of subpersonal use of intentional predicates. Firstly, B&H have extensively documented the use of intentional language in the mind sciences. Secondly, the debate sparked by their work has important diagnostic value. One of B&H's main adversaries is Daniel Dennett, and as I argue in section 5, the exchanges between these parties point to significant weaknesses in the perhaps most prominent theory of subpersonal intentional ascriptions, the intentional stance account. More generally, I take the persistent disagreements and misunderstandings in the debate to nicely illustrate the need for a satisfactory theory of the functioning of intentional concepts in subpersonal explanations. I argue that the topic cannot be dealt with by using conceptual tools from philosophy of mind and philosophy of language alone. Instead, also contributions from the philosophy of science are needed in order to correctly understand the puzzling language use in neuroscience explanations. As a general outcome of the approach developed in the paper, I propose a systematic strategy for answering the question of what kind of language can be used to characterize mechanisms and processes in the mind-brain.

My argument is organized as follows. In section two I present a brief summary of B&H's criticism of cognitive neuroscience. Section three demonstrates how their understanding of what they call the mereological fallacy is flawed, and suggests a more plausible interpretation. I argue that the problem raised by subpersonal applications of intentional language should not be addressed at the semantic level, but that the issue should be reframed as a question concerning the explanatory power of neuroscience theories. A systematic theory of scientific explanation is introduced in section four. The *functional analysis approach* to constitutive explanation shows how two central principles regarding constitutive explanation can be used to assess neuroscientists' contested language use: in sections five and six I argue that although the use of intentional predicates in mechanistic explanations sometimes leads to explanatorily inert claims, intentional predicates can also successfully function as important *heuristic tools* for describing the functional organization of our cognitive architecture.

2 The mereological fallacy argument

Cognitive neuroscience aims to explain people's psychological capacities by revealing their neural underpinnings. These capacities are often referred to by using intentional predicates. Although the concept of intentionality is highly ambiguous, for the present purposes a minimal characterization of the notion will do. Intentional predicates often refer to familiar psychological abilities such as seeing, believing, imagining, guessing and inferring. What can tentatively be said about this diverse group is that they are target-directed capacities (cf. Brentano 1874) paradigmatically manifested by human agents. Moreover, many of the intentional capacities studied in cognitive neuroscience are representational in nature.

However, as B&H point out, intentional concepts are not only used to describe cognitive explananda, but they are also employed in cognitive neuroscience explanations in describing the functioning of the brain and its parts:

> What you see is not what is *really* there; it is what your brain *believes* to be there […] Your brain makes the best interpretation it can according to its previous experience and the limited and ambiguous information provided by your eyes. (Crick 1995, quoted in Bennett & Hacker 2003, 68.)

As the passage from neuroscientist Francis Crick suggests, the use of intentional predicates abounds in the cognitive sciences. It is common to say that the brain believes or infers something, or that one hemisphere has information not accessible to the other, or that the visual system makes guesses about the properties of the distal target of representation. B&H regard this kind of language use as conceptually confused. According to them, the use of intentional predicates in neuroscience explanations amounts to committing *the mereological fallacy*, and results in nonsensical language use.

By 'mereological fallacy' B&H refer to inferences in which the properties of the whole are illegitimately attributed to its parts. The fact that a car is fast does not imply that its carburetor is fast. Analogically, from a person having a belief x one should not infer that the person's brain believes that x.[2] Applications of this argument pattern form the backbone of B&H's attack on several central research programs in cognitive neuroscience: They contest the whole mainstream approach that treats the brain as an information processing system. They also question the meaningfulness of the ideas of there being cognitive maps in the brain, or memories being stored as strengths of synaptic connections (cf. Bennett & Hacker 2003, ch. 3–5). What is common to the criticized research programs is that they involve attributing to the brain cognitive capacities that seem to apply only to whole persons: reading a map, or remembering and handling information.

This concern should be taken seriously. B&H compellingly show that the use of common-sense psychological predicates in mechanistic explanations is a possible source of conceptual confusion, and that the pseudo-problems and illusory understandings created by this confusion might have harmful consequences for scientific practice (Bennett & Hacker 2003, 4–6). As noted above, the main reason for the uneasy reception of B&H's mereological fallacy argument is that it leads to overly strong conclusions. This can be seen by contrasting instances of the fallacy in our folk theory of vision to allegedly fallacious uses of language in contemporary scientific vision research.

As argued by Dennett (1991), our folk theory of vision is susceptible to the idea of the Cartesian theater, which is an illustrative instance of the mereological fallacy. Unlike the folk theory suggests, there is no place in the brain where seeing as such would occur. Instead, shape, color, and movement in the visual scene are all somewhat independently processed by different parts of the visual system. This finding is in line with the mereological fallacy argument: seeing does not seem to be a capacity that a specific part of the brain could have. Importantly, Dennett also claims that modern neuroscience is not entirely free from the Cartesian theater idea either. For example Benjamin Libet's theories and Steven Kosslyn's work on mental imagery fall into the same Cartesian trap as our folk theory (Dennett 2007, 75).

However, a review of the literature suggests that the mereological fallacy argument does not only apply to a few isolated theoretical positions, but generalizes to large parts of contemporary vision research. Cells in the primary visual areas are often said to track orientations in the visual field, and higher visual areas are said to be responsible for movement detection and analysis of complex shapes. The brain is said to use topographic maps in representing visual information. Furthermore, starting from two-dimensional visual cues, the visual system is said to form hypotheses concerning the three-dimensional scene, and to choose between competing hypotheses. (Cf. Wolfe et al. 2009; Chalupa & Werner 2004.) These theories employ intentional language, but to imply their nonsensicality must be misguided. In contradiction to what B&H claim, not only does the intentional language in these explanations make sense, it also seems to play a significant role in augmenting our understanding of the investigated phenomena. Both scientists and philosophers largely agree that these theories offer genuine explanations of how psychological capacities are implemented in the brain. Moreover, the criticized linguistic practices extend beyond vision research and can be found in cognitive neuroscience at large, as well as in other parts of scientific psychology and computer science. Even biological explanations often feature intentional language: in the leading molecular biology journal *Cell*, concepts such as 'controls', 'stimulates', 'silences', 'governs', 'assists', 'orchestrates', 'organizes', 'coordinates', and 'allows' occur among the most common terms referring to activities in biological mechanisms (Moss 2012).

The immediate conclusion from this is that not all ascriptions of intentional predicates to subpersonal entities are equally problematic. There surely are both genuinely fallacious as well as theoretically valuable cases of intentional language in neuroscience explanations. This suggests that the ability to discriminate between the aforementioned cases should be a criterion of adequacy for an account of the use of intentional predicates in neuroscience. The account must offer conceptual resources that make it possible to distinguish fallacious cases from ones in which psychological predicates are fruitfully attributed to subpersonal entities.

3 Dissecting the mereological fallacy argument

The most striking characteristic of the mereological fallacy argument is that it rests on purely semantic considerations. B&H claim that when intentional predicates are attributed to subpersonal entities such as hemispheres, cortical networks, or neurons, concepts are stretched beyond their legitimate scope of application, and thus *nonsense* is produced (Bennett & Hacker 2003, 4–5). They argue that while we understand what it would be for a person to know something, to estimate probabilities, or to form hypotheses, it might be that there simply is no such thing as the brain's thinking or knowing, believing or guessing, or using information (Bennett & Hacker 2003, 70–71).

The immediate reaction to B&H's position would be to dismiss it as a consequence of their refusal to understand the flexibility of natural language. It appears that the ascription of intentional predicates to subpersonal entities should not be taken literally, but that it should rather be understood as metaphorical use of language. B&H do acknowledge this possibility, but claim that their argument still holds: Firstly, it is unclear what the explanatory value of metaphors is (Bennett & Hacker 2008, 249, 257). As I argue in section four, this is a valid concern: scientific explanations ought to describe the causal structure of the world, and it is not obvious how this could be done just by employing metaphors. The main function of metaphor is to illustrate one thing in terms of another, not to explain in any rigorous sense of the word. Furthermore, B&H suggest that even when subpersonal applications of intentional concepts are understood metaphorically, the threat of nonsensicality looms nearby. As they repeatedly claim, meaningfully applying representational language to an entity requires that there is an (i) agent that (ii) uses the representation (symbol, map, and so on) according to a (iii) convention, and that the agent can use it either (iv) correctly or incorrectly. Prima facie, none of these meaning-constitutive criteria apply in the subpersonal case. B&H argue that metaphors are to be judged based on the inferences they license, and since the disanalogy between personal and subpersonal applications of representational concepts is considerable, such metaphors are bound to lead to implausible inferences (Bennett & Hacker 2008, 21, 28, 107, 254–256).

As suggested by John Searle (2007, 107), a natural way to understand this line of argument is to interpret it as an accusation of neuroscientists making a *category mistake* (Ryle 1949/2000). Language games featuring personal-level capacities of human beings differ substantially from subpersonal mechanistic language games, and therefore property attributions transferred from one language game to the other lead to nonsense. This interpretation of B&H's argument is supported by their recurring claim that committing the mereological fallacy leads to nonsensical language use.

Whereas the collapse of meaning is the classic consequence of the category mistake, there is no reason to think that mereological errors should always give rise to nonsense. Thus the most plausible way to understand why B&H claim that mereological fallacies give rise to nonsense is to assume that they equate the mereological fallacy with the category mistake.

Accusing neuroscientists of making a category mistake is certainly a straightforward way to attack cognitive neuroscience. In many ways it is also a very problematic one. As Justin Sytsma (2007) has pointed out, arguments that criticize some applications of a concept of going against its "correct" use rely on problematic presuppositions. For such an argument to get off the ground, one must assume that (i) there exists one well-defined meaning for the concept in question, and that (ii) the person presenting the argument knows this meaning. Furthermore, B&H's insistence that the nature of the fallacy is not empirical but strictly conceptual commits them to (iii) a strong analytic-synthetic distinction. These are highly controversial assumptions regarding the nature of natural language, to say the least. This suggests that the question concerning the role of intentional predicates in mechanistic explanations ought to be raised in a more sophisticated manner.

Although B&H's understanding of the nature of the mereological fallacy is clearly problematic, I suggest that a more plausible interpretation of the issue becomes possible once one acknowledges that category mistake and mereological fallacy are not ultimately the same thing. A mereological fallacy is not fundamentally a mistake in one's language use, but it rather concerns the inferences regarding the relationships between wholes, parts, and their respective properties. Furthermore, it is not always erroneous to ascribe the properties of the whole to its parts. For instance, it would not be mistaken to assume that a handful of sand drawn from a brown pile of sand would also be brown. This applies more generally to so-called aggregative properties, in other words, system properties that are constituted by the properties of its parts alone, not being sensitive to changes in their organization (Wimsatt 2007, ch.12).

Failing to differentiate between category mistake and mereological fallacy leads to conflating two different perspectives, a semantic and an ontological one: On the one hand, the category mistake concerns *how our ordinary mental vocabulary works.* On the other hand, there is the ontological question of *what psychological capacities are*, what they consist of. Unless one is a logical behaviorist about the mental, no drastic conclusions to the ontological question follow from semantic arguments. This is a major weakness of B&H's position: the argument against neuroscience should not be made in the semantic mode. A more useful way to capture the genuine insight behind the mereological fallacy argument is the following. Attributing psychological predicates to subpersonal entities makes neuroscientists vulnerable to charges concerning the explanatory power of their theories. *The properties of a system* (a person's mental capacities) *cannot generally be explained by simply identifying them with the properties of its components* (the brain or its parts). This principle does not follow from semantic considerations, but from the fundamental realist conviction that our explanations must reflect the way the world is.

This alternative way of understanding the mereological fallacy reveals the plausible idea behind many scientists' positive reactions to B&H's worries about language use in neuroscience: linguistic confusions are important because of the implications they could have for the integrity of the explanatory practices of neuroscience. Scientists might not be sensitive to subtle semantic issues as such, but the explanatory power of their theories is a question of paramount importance.

The explanationist framing of the fallacy dovetails nicely with a well-known mereological mistake of psychological explanation, *the homunculus fallacy*. Unlike the category mistake, the homunculus fallacy does not concern language use as such, but it can be understood as a violation of the norms of explanation in scientific psychology. The Cartesian theater fallacy is a fine example of this error. The explanation of our visual capacities should not leave a role for a homunculus, an "inner-seer," in the brain, for this would lead to an infinite regress, and would therefore offer no explanatory gain. Such a postulation would simply be mistaken. Being able to see a visual scene is not the kind of capacity that could be found in some particular part of the brain.

It is time to pause and connect the threads. Neuroscience explanations seem to be guilty of some conceptual violence: when psychological capacities are reductively explained, the explanations often disobey the assumed rules of the language games involving common-sense psychological concepts. It is possible to think that these violations amount to committing a category mistake. However, in some cases, stretching the logic of language accompanies explanatory progress. Whereas our folk theory of vision is simply erroneous in implying the idea of an inner-seer, in scientific vision research statements such as "*the brain makes use of topographic maps*" and "*top-down predictions facilitate the [visual] recognition process by reducing the number of candidate representations of an object that need to be considered*" have played an important role in explaining how our capacity to see is constituted by the workings of our brains (Churchland 2005, 470–472; Kveraga et al. 2007, 149). Therefore, I suggest that in some cases an apparent category mistake is a symptom of the homunculus fallacy, and in others it is not. To systematically distinguish these two kinds of cases from each other one needs a substantial theory of how neuroscience explanation works. In the next section I present the most prominent contemporary theory of psychological explanation, the functional analysis approach. This theory can pin down the conditions under which intentional concepts can successfully feature in mechanistic explanations.

4 The functional analysis approach

Perhaps the most central epistemic goal of cognitive neuroscience is to understand how the behavior of the whole system (the mind) can be explained in terms of the functioning of its components (the brain and its parts). Explanations of this kind are paradigmatic examples of constitutive explanation. As summarized by Carl Craver (2007), there are two main traditions of constitutive explanation in the recent philosophy of science literature: *classical reductionism* and what I will call the *functional analysis approach.*[3]

According to the classical reductionist view (cf. Nagel 1961; Bickle 1998), explanation proceeds by constructing identity statements between the kind-terms of the higher-level theory and those of the lower-level theory, and then deriving the laws of the higher-level theory from the laws of the lower-level theory. Explanation is seen essentially as a matter of subsuming the explanandum event or the higher-level law under the laws of nature. As was argued already by Cummins (1983), this is a problematic view of explanation in the psychological sciences. In the following I rely on an alternative picture of constitutive explanation, the functional analysis approach (Bechtel & Richardson 1993; Bechtel 2008a; Cummins 1983, 2000; Craver 2007). According to this approach, constitutive explanation consists of roughly the following stages:

(1) Tentative identification and description of the explanandum phenomenon.

(2) Recursive functional decomposition of the explanandum phenomenon into its component capacities and their subcapacities.

(3) Localization of these functionally identified components in the actual parts of the physical system.[4]

By showing how the material parts of the system and their organization conspire to constitute the capacities mentioned in the functional decomposition of the explanandum phenomenon, the *explanation discloses how the explanandum phenomenon is brought about by a constitutive mechanism*. In the case of cognitive neuroscience, the explanandum phenomenon is typically a psychological capacity manifested by a person, and explanation proceeds by recursively analyzing this capacity into simpler subcapacities and their organization, ultimately producing a detailed mechanistic description of how the capacities are instantiated by the brain.

As depicted in the Craver diagram in Figure 1, constitutive explanation (as described by the stages 1–3 above) produces hierarchical multi-level descriptions of systems, where a higher-level explanandum capacity (ψ-ing) is explained in terms of the organized functioning of lower-level capacities ($\phi_1, \ldots, \phi_n$), which in turn are explained by decomposing them into yet lower level capacities ($\rho_1, \ldots, \rho_m$).

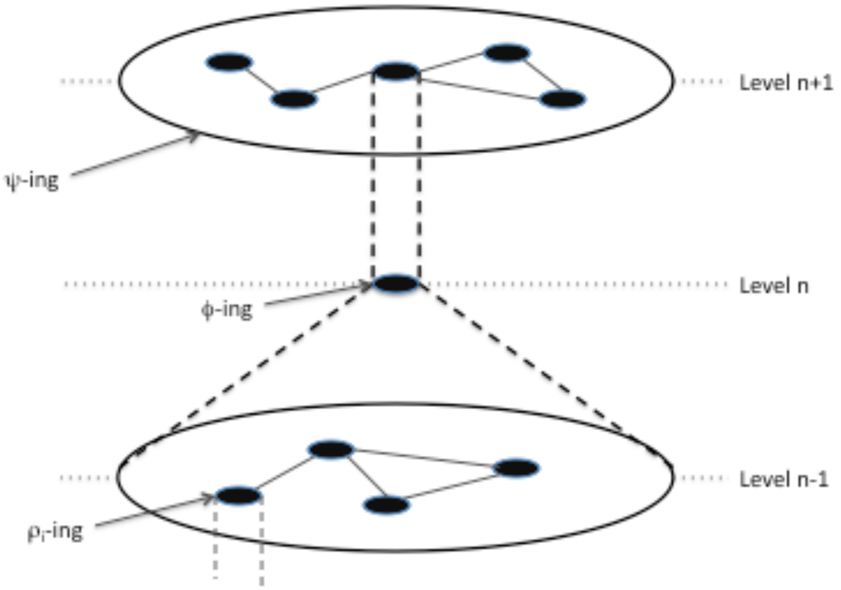

Figure 1. The multi-level organization of a system. Adapted from Craver (2007).

The functional analysis approach has been descriptive by inclination and has primarily aimed at producing realistic characterizations of actual explanatory practices. However, Cummins (2000) and Craver (2007, ch. 4) discuss the often-implicit normative principles that apply to constitutive explanations. These norms can be used to assess the explanatory power of specific explanations. For the present purposes, two fundamental principles can be extracted from the literature:

(FA1) The explanans capacities have to be simpler than the explanandum capacity.

The explanatory interest of functional analysis is proportional to the extent to which the explanans capacities are less sophisticated than the explanandum capacities. This is because the explanatory gain comes from showing how the simpler capacities are orchestrated together to produce the more complex explanandum phenomenon (Cummins 2000). In mechanistic terms, appealing to the organization between components is the key to showing how complex system capacities can emerge from simpler activities exhibited by the components of the mechanism.

(FA2) The subcapacities used to decompose the explanandum must eventually be localized in concrete causal structures.

Explanations need to be descriptions of causal mechanisms. For an explanation to be acceptable, one must be justified in one's belief that the capacity is realized by an actual causal structure in the brain (Craver 2007, 115). *FA2* shows why the explanations produced by the functional analysis approach can be conceived as mechanistic explanations. While it is usually possible to functionally decompose systems in a number of different ways, only those decompositions whose component capacities can be anchored in the functioning of the actual structural parts of the system (stage 3) can be considered as genuine mechanistic causal explanations.

5 The heuristic role of intentional predicates in subpersonal explanations

Unlike in the classical reductionist view, in the functional analysis approach the relationship between mental phenomena and their neural basis is not understood as a mapping between two levels, the mental and the physical, but rather conceived as a hierarchy of abstract descriptions of the same system. The transition from the "two-levelism" of the classical view (Lycan 1990) to this hierarchical picture makes it possible for intentional predicates to appear below the personal level: I suggest that the use of intentional language at subpersonal levels must be understood as a tool for the functional description of subcapacities. *Common-sense psychological terms can be used to functionally decompose a higher-level explanandum into its subcapacities* (stage 2 of the explanatory process, see above). That is, in the context of cognitive neuroscience, intentional language should be understood as an epistemic tool for describing the abstract functional roles played by the components of the cognitive system.

To illustrate how intentional predicates can be applied to subpersonal entities, let us consider the example of how the predicate 'to guess' could be used in describing the functioning of the human visual system. What we ordinarily mean by someone guessing the answer to a question is that (i) based on the given partial information, (ii) relevant background assumptions, and a (iii) set of alternative answers, the person (iv) (inductively) chooses the best available answer. This can be thought of as a rough characterization of the minimal criteria for the application of the concept.[5] In a similar way, when applied to visual recognition, guessing could be understood to imply the following. Relying on (i) ambiguous two-dimensional visual cues and (ii) either learned or hard-wired assumptions built into the visual system (about the direction of light, etc.) as inputs, the visual system chooses between (iii) competing interpretations of the data, resulting in often successful (iv) recognition of the distal target. On a relatively abstract level of description, as in the present case, the functional relationship between the inputs and outputs of a cognitive capacity can thus be captured by a familiar psychological term ('to guess'), and the intentional predicate can be a succinct way to tentatively express the functional role played by a component of the visual system. Importantly, if this causal role can also eventually be described in terms of lower-level mechanisms, the use of intentional language need not undermine the explanatory power of the theory.

Despite its simplicity, this analysis of guessing arguably connects to prominent developments in vision research extending from early 19th century psychophysics to contemporary computational modeling of visual recognition. Already Hermann von Helmholtz (1866) proposed that perception is not an exclusively stimulus-driven process, but that the brain also utilizes knowledge of the past to infer the properties of external objects and events. In the 1990s, this fruitful but perhaps elusive picture of the brain as a prediction device got a more precise mathematical expression in the theories of the Bayesian brain paradigm (cf. Knill & Pouget 2004). In this modeling approach, visual recognition is approached as a probabilistic inference process, in which incoming sensory data functions as accumulating evidence for testing competing top-down hypotheses about distal targets of perception. The mechanistic plausibility of this functional-decomposition strategy is currently under careful scrutiny, but although the details of neurally tractable mechanisms capable of implementing the required probabilistic guessing or predicting are still open, it is undeniable that characterizing the brain as a prediction machine has been a very fruitful hypothesis. It has resulted in formulation of powerful computational models and it guides the search for the corresponding neural mechanisms. (Cf. Friston 2011.)

This brief example suggests that intentional language can often play a crucial *heuristic role* in cognitive neuroscience explanations. At the early phases of theory formation, when there is little information about the mechanisms and processes behind the phenomenon, intentional predicates can feature in the preliminary analysis of the explanandum by suggesting possible functional decompositions of the phenomenon. This might often be the case in many research programs within young fields like contemporary cognitive neuroscience: Although on the one hand there is robust scientific evidence concerning the low-level cellular mechanisms, and on the other hand plenty of brain-imaging data of high-level processes, the experimental evidence concerning the crucial middle-level mechanisms, cellular networks, is scarce. I suggest that *intentional language is often needed for filling in the gappy, multi-level theories of cognitive capacities:* in the somewhat uncharted territory of middle-level cognitive architecture, intentional concepts can be important tools for formulating hypotheses. At later stages of theorizing models are usually expressed in computational terms, but intentional characterizations can provide the required first sketch – computational models do not usually come out of the thin air, but instead often reflect some prior understanding of what happens in the system. Intentional predicates can be used in characterizing the capacities in middle-level mechanisms in an understandable way, and thus they have the potential to create understanding not achieved by other available methods.

6 How to describe mechanisms

The above considerations suggest that the explanatory role of intentional concepts should not be downplayed as "merely" heuristic. Heuristic tools often play a crucial part in scientific research processes (Wimsatt 2007). However, this does not mean that mechanistic explanations in cognitive neuroscience always need to involve intentional language.[6] As stated by *FA2*, descriptions of causal mechanisms must eventually bear all the explanatory weight in cognitive neuroscience theories. Therefore, the crucial issue is *how the causal mechanisms investigated in cognitive neuroscience ought to be described*. The question is far from trivial. As pointed out earlier, according to the functional analysis approach, explanations of cognitive capacities require describing them at various levels of analysis. Although there is no sharp dividing line between levels that are acceptably described in representationalist terms and "purely" mechanistic ones (Lycan 1990), it seems likely that as one goes down in the hierarchy of decomposition of capacities, one finds fewer and fewer components that could reasonably be described with intentional predicates. That is, whereas on the higher levels of a cognitive system one can find complex capacities that are somewhat similar to the capacities of persons (e.g., 'to guess'), the functional organization on the lower levels might be such that its elements cannot accurately be described in familiar psychological terms. The functional roles at these levels are unlikely to be *semantically transparent* (Clark 1989), and should therefore be described in alternative ways, often perhaps minimally in terms of input-output-patterns.

Moreover, *FA1* and *FA2* together imply that components and activities featuring in cognitive neuroscience explanations must always be compatible with descriptions of the functioning of cellular-level structures. The cellular level appears to be fundamental for cognitive neuroscience in the sense that although the entities and activities at that level can be further decomposed into simpler biochemical structures, the information gained from these analyses often falls outside the explanatory interests of cognitive neuroscience. Therefore, the level of individual neurons and their activities can plausibly be thought to be the default level at which cognitive neuroscience explanations bottom out. In consequence, it might seem troubling that even at this level the capacities can sometimes be described in representational as well as in structural terms. The behavior of a V1-neuron can be described in terms of membrane potentials, spike trains, and other non-representational cellular activities, but the cell can also be said to track features or spatial frequencies in the visual array.[7] However, the functional analysis approach suggests that on the lowest level of a mechanistic explanation, the functional components must be anchored in actual causal structures by way of a strict coincidence between the functional and structural descriptions. That is, to genuinely localize the low-level capacities in cellular structures, there should be no "free-floating" functional postulations, but for all capacities at this level it should be possible to show how they correspond to a structural description of cellular level happenings that can be unanimously interpreted as carrying out the execution of the capacity (cf. Cummins 1989, 88–90). Here the explanatory promises made at higher levels by employing functional characterizations (e.g., intentional predicates) must be redeemed by handing in structural descriptions.

This mechanistic constraint can be used to make explicit the distinction between heuristic and genuinely explanatory descriptions of explananda. Heuristically useful intentional predicates can produce potential understanding by suggesting theories of the functional organization of the system in question. These suggested decompositions can be called *how-possibly explanations* (Craver 2007, 112). Yet if later in research it turns out that the functional decomposition achieved by using intentional predicates does not capture anything describable in lower-level structural terms, one has to admit that the understanding given by the suggested decomposition was illusory. However illuminating, unless it is a correct description of the actual causal structure of the system, it fails to qualify as a *how-actually explanation*. Unlike how-possibly explanations, how-actually explanations not only tell how things could be, they tell how things actually are, and thus qualify as genuine scientific explanations of their targets' properties. In sum, according to my account, there is no conceptual reason prohibiting the use of intentional predicates in subpersonal mechanism descriptions, but their use has to be justified by compatibility with lower-level accounts of the functioning of the same system.

Above it was argued that intentional predicates might not often be the conceptual resources favored in low-level neuroscience explanations. Also on higher levels of the constitutive hierarchy, there are good reasons to replace intentional descriptions of capacities with technical scientific representations as research progresses. This is a somewhat idealistic but still reasonable aim, because there is no a priori reason for why capacities on all subpersonal levels of explanation could not *in principle* be described without employing intentional vocabulary. The main reason for preferring computational and algorithmic theories to ones employing intentional vocabulary is that the characterizations of capacities given by using intentional predicates are bound to be vague, and will not usually pick out exactly the same causal structures as more rigorous descriptions of the same function. This is because the meanings of our common-sense psychological concepts are very rich. Many of their everyday connotations do not apply at the subpersonal level, and therefore below the personal level they must be used in an impoverished way. One either has to specify the intended meaning of the concepts beforehand (as in the example above), or leave the intended interpretation open. The former strategy seems to make the intentional predicate itself redundant, and the latter one leaves enough room for interpretation for the conceptual confusions anticipated by B&H to arise.

This picture of constitutive mechanistic explanation finally allows us to more precisely describe the nature of the homunculus fallacy. The crux of the fallacy seems to be the *explanatory inertness* of the fallacious claim, not semantic confusion as suggested by B&H. Postulating an undischarged homunculus (an inner-seer, for example) fails to explain because merely transporting a personal level property down to a lower level does not do any explanatory work (*FA1*). In the case of the visual homunculus, the norm concerning localization (*FA2*) is obviously also violated. There is no identifiable structure in the brain implementing the role of the inner-seer; the visual homunculus is simply nowhere to be found in the brain. In contrast, contemporary vision research is an example of highly successful research program within which intentional language has often been employed at subpersonal levels. It explains vision by dividing the explanandum capacity (the ability to see) into several sub-tasks (shape, color, and movement processing), and each of these sub-tasks is again subdivided into further subcapacities, which are ultimately localized in neural structures and processes. The research on the early stages of retinal and cortical visual processing is an ideal case of this process. Even the cellular-level causal mechanisms at the early stages are known particularly well, and thus the subcapacities featuring in the explanations have been satisfactorily localized in well-delineated brain structures (cf. Kalat 2004, ch. 6).[8]

7 Coda: To explain is not to eliminate

In this penultimate section I bring my account of the explanatory role of intentional predicates to bear on the persistent disagreement between B&H and their most prominent opponent Daniel Dennett. As suggested above, I hold that the stagnant state of the debate results from both parties lacking a crucial conceptual resource, a plausible theory of explanation for the mind sciences.

Although there are important similarities between Dennett's intentional stance approach and the functional analysis theory of constitutive explanation, Dennett's disregard for the mechanistic constraints in explanation implies that his theory does not qualify as a full-blown theory of causal explanation. According to Dennett, the explanatory strategy of the intentional stance consists in analyzing rational behavior into sub-tasks that themselves can be described as if they were rational. This justifies the use of intentional language at subpersonal levels (Dennett 1987). Computers, brains and even brain cells are systems that can have psychological properties in an attenuated sense (Dennett 2007, 87–8). Although Dennett recognizes that the homunculi thus produced must be discharged by further analyzing them into less clever homunculi, he holds that subpersonal entities are *sufficiently similar* to persons to merit intentional characterization.

This as-if language has often raised justified accusations of instrumentalism against Dennett, because it invites one to understand the subpersonal ascriptions as merely metaphorical use of concepts, not as a literal description of what happens in the cognitive system in question. As was argued above, it is not clear what explanatory work such metaphors could do. For instance, adopting the intentional stance and claiming that a thermostat "knows" the temperature clearly makes sense and might be sufficient for predicting its functioning under normal circumstances, but obviously does not qualify as a scientific explanation. I suggest that intentional stance is better understood as a strategy for creating novel decompositions of complex phenomena: it is an interpretative perspective that recursively treats systems and their subsystems as goal-directed and rational, and can thus illuminate patterns in their behavior that perhaps would not be visible otherwise. What makes the stance so powerful is that it can gloss over major mechanistic differences in the architectures of different kinds of cognitive systems. However, as has already been emphasized, merely suggesting different functional decompositions of a psychological capacity does not amount to giving causal explanations. At best, these suggestions are *how-possibly explanations* of functional organization, hypotheses to be examined by empirical research on causal mechanisms. Failing to pay attention to the importance of localization (stage 3 of the explanatory process), Dennett's stance stance cannot give an adequate picture of the functioning of intentional predicates in neuroscience explanations, and therefore the analysis of the explanatory use of intentional predicates in neuroscience is better conducted in light of the functional analysis approach.

On the other hand, B&H's confusions regarding scientific explanation stem from two sources. Firstly, the little that they say about the topic suggests that they adhere to the classical reductionist view of scientific explanation (Bennett & Hacker 2003, 362–3). Therefore, many of their arguments against explanation in cognitive neuroscience are directed at this rather outdated target and do not apply to the functional analysis approach. Secondly, B&H are generally skeptical of whether it is at all possible to illuminate the nature of personal-level psychological phenomena by examining their neural basis (Bennett & Hacker 2003, 3). This strongly anti-reductionist attitude is inspired by certain strands of Wittgensteinian philosophical psychology. Ultimately the worry appears to be a semantic variant of the mind-body problem: Going inside a brain one would not see thoughts but just mechanical operations of cellular cogs and wheels. For instance, mere functioning of mechanisms does not exhaust the phenomenon of being able to see something. Quite the opposite; the content and truth conditions of personal-level concepts differ from those of subpersonal concepts. Intentional phenomena cannot simply be *replaced* by non-intentional neural mechanisms. It would seem that instead of really explaining psychological capacities, reductive explanations amount to merely changing the topic.

As I argue below, I think that the argument relies on a misunderstanding concerning the nature of constitutive explanation. However, it does spot a widespread error: explaining is often conflated with making straightforward identification claims. As B&H painstakingly document, even prominent neuroscientists sometimes express views according to which the nature and the behavior of psychological entities are to be *completely* explained in terms of their parts, and that the entities *are* nothing but their parts (Bennett & Hacker 2003, ch. 13). B&H have not been the only ones to pay attention to the reductionism prevailing in many parts of neuroscience. In his defense of ruthless reductionism, John Bickle (2003, 2006) has alluded to mainstream work in neuroscience as a proof of the plausibility of his strongly reductionist account, and in a recent essay Michael Gazzaniga (2010) explores the reasons behind neuroscientists' adamant reductionist tendencies. Many of the views discussed in these sources amount to strong ontological reductionism, a problematic position rightly criticized by B&H. However, the functional analysis approach explored in this paper does not carry such eliminativist commitments. Explaining does not mean explaining away.

To grasp the distinction between explanatory and ontological reductionism, one must understand what successful constitutive explanations can accomplish. A constitutive explanation offers *information on very specific aspects* of the explanandum phenomenon. By showing how the properties and the functioning of the explanandum capacity depend on the states and the organization of its subpersonal components, an explanation gives contrafactual information on how the whole would change as a result of changes in its components and their organization. For example, by knowing the functional organization of the visual system, one could answer questions concerning how damage in a specific brain area would affect one's visual abilities. Consequently, once constitutive explanation is understood in this way, psychological phenomena do not present a special a-priori problem for constitutive explanation. Claiming that mechanistic explanations fail to illuminate mental explananda is ultimately on a par with the claim that referring to the functioning of the cogs and wheels inside a grandmother clock fails to explain the clock's capacity to tell the time.

In addition to the ontological non-reducibility argument, B&H argue against neuroscience explanations by drawing attention to their limited scope. They point out that many psychological phenomena involving reasons, actions, social norms, and institutions cannot be explained with neuroscientific resources (Bennett & Hacker 2003, 360–5). This argument builds on the idea that whereas personal-level explanations are inherently normative, no normativity appears in "brute" causal explanations at subpersonal levels. Personal- and subpersonal-level explanations are different in kind, and therefore incompatible. This position against subpersonal explanation has been forcefully defended by several authors (cf. McDowell 1994; Hornsby 2000). Without delving deeper into the controversial topic, it is sufficient to restate the idea from the previous paragraph: Like all other kinds of explanation, constitutive explanation can only answer a limited range of questions, and is thus aspectual. It does not promise an exhaustive characterization of all the dimensions of the explanandum phenomenon. There are a variety of questions not satisfactorily answered with reductive explanations. Luckily the relationship between different kinds of explanations does not need to be as antagonistic as has been thought. Rather than exclusive, different explanations should often be seen as complementary.

8 Conclusions. From grammatical errors to inert explanations

In this paper I have developed an account of the functioning of intentional predicates in cognitive neuroscience theories. Subpersonal mechanistic explanations employing intentional language raise a genuine problem worth philosophical scrutiny, but I have argued that simply looking at the linguistic form of scientists' assertions is not sufficient for distinguishing confused uses of language from theoretically valuable ones. In order to move beyond the crude strategies of indiscriminate rejection or acceptance of the controversial language use, one must abandon the purely semantic approach and take into account the explanatory context. Unconventional use of intentional concepts should be seen as a linguistic reflection of the explanatory process.

Two fundamental normative principles regarding constitutive explanation were extracted from the functional analysis approach. They can be used to distinguish fruitful uses of intentional predicates from instances of the homunculus fallacy. According to my explanationist approach, genuine homunculus fallacies violate the norms of constitutive explanation, whereas intentional predicates can also acceptably function as heuristic tools for decomposing the explanandum phenomenon into its functional components. This framing of the mereological fallacy argument helps make visible the practical implications of the issue: the problem that sometimes arises with the use of intentional language in mechanistic explanations is not the descent into nonsense, but rather the explanatory inertness of the fallacious claim.

Ultimately the problem dealt with in this paper is a part of a larger question concerning mechanistic explanation: How can mechanisms in general be characterized? Which part of our conceptual toolbox is applicable to describing causal mechanisms? The answer seems to vary between the different sciences. The fundamental activities in biological mechanisms are simpler than the more abstract capacities and activities featuring in psychological mechanisms. Thus the question clearly cannot be given a single correct answer. However, abandoning traditional two-levelism in favor of the functional analysis approach suggests a promising strategy for attacking this more general question.

The strategy comes in three stages. In order to judge whether a concepts is applicable in a certain explanatory context, one should start by first clarifying the intended meaning of the concept in that particular context, and thereby determine the inferential commitments one makes by using it. This makes explicit the functional role that the concept individuates. Secondly, different scientific disciplines deal with different levels of the hierarchical structure of complex systems. The higher the level of mechanism, the more likely the happenings there are to be well suited for description with inferentially rich concepts. That is, whereas the fundamental capacities in biological mechanisms tend to be quite mechanical in the classical sense of the word (e.g., ion channels opening, helices unwinding), at the levels investigated by cognitive psychology the activities concern information processing and can be captured by using more complex, often representational, concepts (cf. Darden 2008; Bechtel 2008b). Locating the investigated mechanism at a certain level in the explanatory hierarchy is thus the second step in determining what concepts can be used for describing it. Finally, there is a further level-relative constraint resulting from the explanatory asymmetry between the higher and lower levels: higher-level capacities featuring in explanations have to be ultimately cashed out in terms of lower-level descriptions. Therefore, it should be ascertained that inferential commitments made in characterizing higher levels are compatible with lower-level descriptions of the system. This inferential bookkeeping strategy results in *cautious liberalism about mechanism description*: integrative multi-level explanations in the cognitive sciences can employ a variety of representational tools, as long as it is possible to meet the conceptual goal of explicating the relationships between different (intentional, computational, and structural) descriptions and thus interweave the different representational elements into a coherent whole.

Notes

[1] B&H claim that, among several others, the work researchers such as Francis Crick, Antonio Damasio, Michael Gazzaniga, Richard Gregory, Eric Kandel, and David Marr builds on conceptually confused foundations.

[2] B&H are not the first ones to present this criticism of scientific psychology. They often quote Wittgenstein's dictum, according to which only human beings could meaningfully be said to sense, see, hear, or to be conscious (Wittgenstein 1958, §281; see also Kenny 1971; Chisholm 1991).

[3] The functional analysis approach (Cummins 1983) goes also by several other names such as 'the systems tradition' (Craver 2007) and 'heuristics of decomposition and localization' (Bechtel & Richardson 1993). Despite their differences, these views have enough in common to allow common exposition.

[4] This is a highly simplified picture of the explanatory process. In real research, stages 1-3 do not need to temporally follow each other. Resulting from bottom-up research strategies, the mechanistic decomposition (stage 3) often considerably influences the identification of the explanandum phenomenon and its functional decomposition (cf. Bechtel 2008a). Nor is this characterization meant to suggest that a convenient form-function correlation between the physical structure and the functional decomposition could always be found; often evolved cognitive systems show breakdowns of clear modularity required by neat form-function correspondence.

[5] This "well-informed guessing" is of course only one use of the term, and there are several others. Guessing under conditions of complete ignorance, such as choosing Lotto numbers, would be one such alternative meaning for the word.

[6] Except, or course, in describing psychological explanandum phenomena (Clark 1989, 50–52).

[7] For examples of the difficulty of finding the correct functional descriptions of single-cell functioning, see Uithol et al.'s (2011) discussion of the mirror neuron system.

[8] For an ambitious agenda for banishing the homunculus from theoretical vision research, see Barlow 1995.

References

Barlow, Horace. (1995). Banishing the Homunculus. In D. Knill and W. Richards (Eds.) *Perception as Bayesian Inference.* Cambridge: Cambridge University Press.

Bechtel, William. (2008a). *Mental Mechanisms*. New York: Routledge.

Bechtel, William. (2008b). Mechanism in Cognitive Psychology: What are the operations. *Philosophy of Science* 75, 983–994.

Bechtel, William, and Robert Richardson. (1993). *Discovering Complexity*. Princeton: Princeton University Press.

Bechtel, William & Adele Abrahamsen (2005). Explanation: a mechanist alternative, *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 421–441.

Bennett, Maxwell, and Peter Hacker. (2003). *The Philosophical Foundations of Neuroscience*. Malden: Blackwell.

Bennett, Maxwell, and Peter Hacker. (2008). *History of Cognitive Neuroscience*. Wiley-Blackwell.

Bennett, Maxwell, Daniel Dennett, Peter Hacker, John Searle. (2007). *Neuroscience and Philosophy: Brain, Mind, and Language*. New York: Columbia University Press.

Bickle, John. (1998). *Psychoneural Reduction: The new wave*. MIT Press.

Bickle, John. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Springer-Verlag, New York.

Bickle, John. (2006). Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience, *Synthese*, 151, 411-434.

Brentano, Franz (1874) *Psychology from an Empirical Standpoint*, London: Routledge and Kegan Paul.

Chalupa, Leo M., and John S. Werner. (ed.) (2004). *The Visual Neurosciences*. Cambridge: MIT Press.

Chisholm, Roderic. (1991). On the simplicity of the soul. *Philosophical Perspectives* 5, 167–180.

Churchland, Paul (2005). Cleansing science. *Inquiry* 48, 464–477.

Clark, Andy. (1989). *Microcognition. Philosophy, Cognitive Science and Parallel Distributed Processing*. MIT Press.

Craver, Carl. (2007). *Explaining the Brain.* Oxford: Clarendon Press.

Crick, Francis. (1995). *The Astonishing Hypothesis*. London: Touchstone.

Cummins, Robert. (1983*). The Nature of Psychological Explanation*. Cambridge: MIT Press.

Cummins, Robert (1989). *Meaning and Mental Representation*. Cambridge: MIT Press.

Cummins, Robert. (2000). 'How does it work?' versus 'what are the laws?': Two conceptions of psychological explanation. In *Explanation and Cognition*, ed. Frank Keil, and Robert Wilson, 117–144. MIT Press.

Dainton, Barry. (2007). Wittgenstein and the brain. *Science* 317, 901.

Darden, Lindley. (2008). Thinking Again about Biological Mechanisms. *Philosophy of Science* 75, 958–969.

Dennett, Daniel. (1987). True Believers. The intentional strategy and why it works. In Dennett, D. *The Intentional Stance*, 13–42. MIT Press: Cambridge.

Dennett, Daniel. (1991). *Consciousness explained*. Penguin books.

Dennett, Daniel. (2007). Philosophy as naïve anthropology: Comment on Bennett and Hacker. In *Neuroscience and Philosophy*, ed. Bennett et al., 73–96. New York: Columbia University Press.

Friston, Karl. (2011). The history of the future of the Bayesian brain, *NeuroImage*, corrected proof, Oct 2011.

Gazzaniga, Michael (2010). Neuroscience and the correct level of explanation for understanding mind, *Trends in Cognitive Sciences*, 14, 297–292.

Glennan, Stuart. (2002). Rethinking mechanistic explanation*, Philosophy of Science*, 69, 3 (supplement).

von Helmholtz, Hermann. (1866). Concerning the perceptions in general, 3rd ed. *Treatise on Physiological Optics, Vol. III* (translated by J. P. C. Southall 1925 Optical Society of America. Section 26, reprinted New York: Dover, 1962).

Hornsby, Jennifer. (2000). Personal and sub-personal; A defence of Dennett's early distinction. *Philosophical Explorations* 3, 6–24.

Kalat, James. (2004). *Biological Psychology*. 8th edition. Thomson Wadsworth.

Kenny, Anthony. (1971). The homunculus fallacy. In *Interpretations of life and mind: Essays around the problem of reduction*, ed. Marjorie Grene, 65–74. London: Routledge.

Kenny, Anthony. (2008). Foreword. In *History of Cognitive Neuroscience*, Maxwell Bennett, and Peter Hacker, xvii–xix. Wiley-Blackwell.

Knill , David, and Alexandre Pouget. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation, *Trends in Neurosciences*, 27, 712–719.

Kveraga, Kestutis, Avniel Ghuman, and Moshe Bar. (2007). Top-down predictions in the cognitive brain, *Brain and Cognition,* 65, 145–168.

Lycan, William. (1990). The continuity of levels in nature. In *Mind and Cognition. An anthology,* ed. William Lycan, 77–97. Oxford: Wiley-Blackwell.

Machamer, Peter (2004). Activities and causation: The metaphysics and epistemology of mechanisms, *International Studies in the Philosophy of Science*, 18, 27–39.

Machamer, Peter, Lindley Darden, and Carl Craver (2000). "Thinking about mechanisms", *Philosophy of Science* 57, 1–25.

Machamer, Peter, and Justin Systma. (2009). Philosophy and the Brain Sciences. *Iris: European Journal of Philosophy and Public Discourse* 1, 65–86.

McDowell, John. (1994). *Mind and World*. Cambridge: Harvard University Press.

Moss, Lenny (2012). Is the philosophy of mechanism philosophy enough? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences,* 43, 164-172.

Nagel, Ernst. (1961). *The Structure of Science*. Harcourt, Brace & World.

Patterson, Dennis. (2003). Review of Philosophical Foundations of Neuroscience. *Notre Dame Philosphical Reviews*. http://ndpr.nd.edu/review.cfm?id=1335. Accessed 11 January 2011.

Ryle, Gilbert. (1949/2000). *The Concept of Mind*. University of Chicago Press.

Searle, John. (2007). Putting consciousness back in the brain: Reply to Bennett and Hacker", In *Neuroscience and Philosophy*, ed. Bennett et al., 97–126. New York: Columbia University Press.

Schaal, David. (2005). Naming our concerns about neuroscience: A review of Bennett and Hacker's Philosophical Foundations of Neuroscience. *Journal of the Experimental Analysis of Behavior* 84, 683–692.

Slaney, Kathleen, and Michael D. Maraun. (2005). Analogy and metaphor running amok: An examination of the use of explanatory devices in neuroscience. *Journal of Theoretical and Philosophical Psychology* 25, 153–173.

Sytsma, Justin. (2007). Language Police Running Amok. *Journal of Theoretical and Philosophical Psychology* 27, 89–103.

Uithol, Sebo, Iris van Rooij, Harold Bekkering and Pim Haselager (2011). What do mirror neurons mirror? *Philosophical Psychology*, 24(5), 607–623.

Wimsatt, William. (2007). *Re-engineering Philosophy for Limited Beings*. Cambridge: Harvard University Press.

Wittgenstein, Ludwig. (1958). *Philosophical Investigations* 3rd edition. Upper Saddle River: Prentice Hall.

Wolfe, Jeremy M. (et al.). (2009). *Sensation and Perception*. Sunderland: Sinauer Associates.