Thoughts, Motor Actions, and the Self

Article *in* Mind & Language · February 2007

DOI: 10.1111/j.1468-0017.2006.00298.x

CITATIONS
61

READS
253

2 authors:

Gottfried Vosgerau
Heinrich-Heine-Universität Düsseldorf
77 PUBLICATIONS **1,717** CITATIONS

SEE PROFILE

Albert Newen
Ruhr-Universität Bochum
115 PUBLICATIONS **4,596** CITATIONS

SEE PROFILE

Thoughts, Motor Actions, and the Self

GOTTFRIED VOSGERAU AND ALBERT NEWEN

Abstract: The comparator-model, originally developed to explain motor action, has recently been invoked to explain several aspects of the self. However, in the first place it may not be used to explain a basic self-world distinction because it presupposes one. Our alternative account is based on specific systematic covariation between action and perception. Secondly, the comparator model cannot explain the feeling of ownership of thoughts. We argue—contra Frith and Campbell—that thoughts are not motor processes and therefore cannot be described by the comparator-model. Rather, thoughts can be the triggering cause (intention) for actions. An alternative framework for the explanation of thought insertion in schizophrenics is presented.

1. The Comparator-Model

In the 1950s von Holst and Mittelstaedt (1950) and Sperry (1950) developed the comparator-model. It explains and describes motor control mechanisms in living things. The starting point was the fact that classical reflex theory had difficulty describing the different behaviour as response to the same stimulus depending on whether the stimulus is caused externally or internally (by motion of the animal). A fly, for example, sitting in a vertically striped cylinder will turn if the cylinder turns around it. The movement of the fly is such that the change in visual flow is nullified (optomotoric reflex). If, however, the fly turns itself while the environment is stable and thereby causes the same change in the visual flow, it will not show this optomotoric reflex. The fact that reactions depend on the cause of the stimulus cannot be explained by classical reflex theory alone.

The core idea of the 'reafference principle' (von Holst and Mittelstaedt, 1950) is that the input stimulus (afference) does not directly trigger the response. If a purposeful action is executed, not only does the efferent signal (the so-called efference) activate the muscles but also an efference copy is made. This efference copy effectively nullifies the so-called reafference, i.e. the afference that is caused by the movement. In terms of action potentials, the efference copy is the inverse

We are grateful to the *VolkswagenStiftung* for their kind financial support of the research project 'Self-Consciousness and Concept Formation in Humans' under the supervision of Professor Newen which made the research for this article possible. For critical discussion and helpful comments we would like to thank John Campbell, Vera Hoffmann, Jan Restat, Matthis Synofzik, Kai Vogeley, Alexandra Zinck, and two anonymous reviewers.

Address for correspondence: Gottfried Vosgerau, Philosophisches Seminar, Bursagasse 1, 72070 Tübingen, Germany.
Email: vosgerau@uni-tuebingen.de

Mind & Language, Vol. 22 No. 1 February 2007, pp. 22–43.

of the reafference such that the juxtaposition of both results in zero, hence no activation (see Figure 1).[1] In general, the efference copy and the afference (including reafference) are compared, i.e. the difference between them is computed. This difference between efference and afference is exactly the part of the afference that is not *re*afference, i.e. the part that is not caused by the system itself. Only this signal is further processed. There are two special cases: First, if a motor command comes from a higher centre and no external change (which is not caused by the acting system) takes place, the afference will consist of the reafference alone. In this case, the difference between efference copy and afference will be zero and no reflex will be triggered. Second, when there is no motor command and hence the efference copy is zero, the difference will equal the afference, i.e. every external change is further processed. In all other cases the comparator 'filters' the incoming signal (the afference) by computing the self-caused part (reafference) from the input, such that only externally caused afferences are forwarded to higher systems.

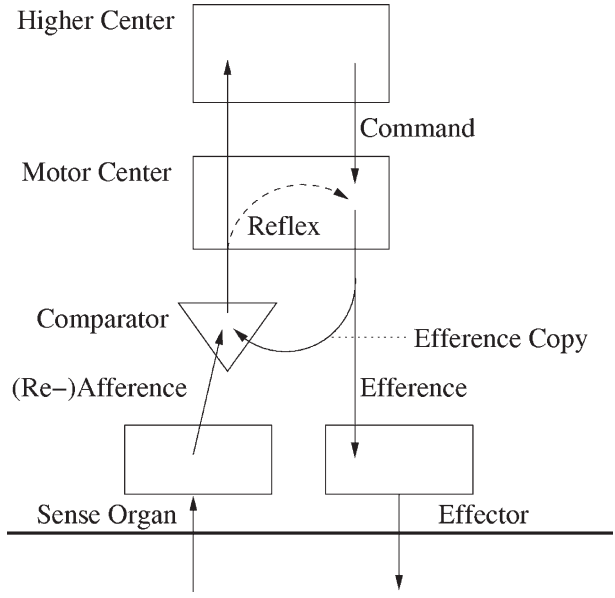

Figure 1 *The reafference model*

[1] von Holst and Mittelstaedt (1950) speak of inverse activations (labelling it with '+' and '-'), whereas Frith (1992) remains quite vague of how the comparator exactly works. Grush (2004) recently proposed a general model that also comprises comparators: The comparison made by the 'Kalman Filter' is exactly the computation of the difference between efference copy and afference ($I^{*\prime}(t) - S(t)$ in his terms, p. 380). We think, that every story of comparators has to come down to this picture. We will discuss this point in more detail in section 2.1.

More generally, the reafference model can be characterized by a comparator that compares the efference copy and the afference (see Figure 1). The difference between the two is transferred to the motor centre (and to higher centres). In the motor centre it can directly trigger a reflex. The reflex is suppressed, however, when there is a motor command, since in this case the efference copy nullifies the afference. This means that in such cases the afference is not transferred to higher centres either. In fact, as Helmholtz (1866) already noticed, the world does not seem to move when we move our eyes. The efference copy of the eye movement command nullifies the movement of the picture on the retina. However, if we push our eye with a finger, the world really seems to make a jump, for the efference copy of this finger movement is not used to 'correct' the visual input since finger movements do not usually affect the retinal picture.

Frith (1992) extended this model to explain certain symptoms of schizophrenia, especially delusion of control and thought insertion. In Frith, Blakemore and Wolpert (2000) he proposes a model of motor control that accounts for the explanation of several abnormalities in awareness and control of action. For human motor control, he proposes a self-monitoring system that is composed of three separate comparators (see Figure 2).

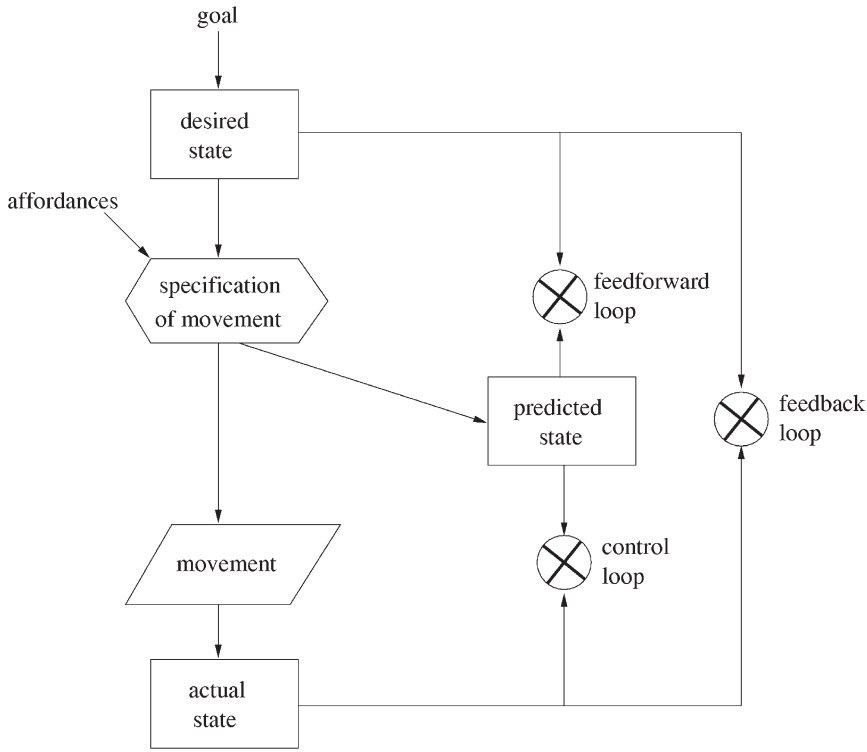

Figure 2 *Monitoring systems (cf. Frith, Blakemore and Wolpert, 2000, p. 1784)*

First, the desired state is compared with the predicted state in order to control the specification of movements to be made (feedforward loop). The outcome can also be used to improve the specification with mental practice. Second, the actual state is compared with the predicted state in order to control and eventually adjust the performed movements (control loop). Improvement of the prediction is based on this comparator. Third, the actual state is compared with the desired state in order to control the success of the action (feedback loop), and thereby allows improvement of the whole system of motor command generation.

Delusion of control can be explained by an impairment of the prediction. If there is no predicted state, there is neither a match between the desired and the predicted state nor between the predicted and the actual state. Therefore, patients are not aware of initiating and controlling the action, nor do they know about the consequences of their action. However, because the desired state still matches the actual state, they recognize that their action conforms to their intention. This results in the feeling that their intentions are 'read' by somebody who then makes the movement for them.

In his (1992) book, Frith applies this model to thought insertion. However, he assumes that thoughts can be analyzed as inner speech and hence as motor processes. Just as in the case of motor control, a match between the desired and the predicted state could be used to label the thought as produced by me. If the prediction mechanism is impaired, thoughts can occur that do not receive that label and are therefore experienced as coming from an external source. Nevertheless, it does not seem accurate to explain thinking by inner speech. First, the inner speech model can capture at most occurent thoughts; all the background knowledge and beliefs that play a role in thinking and decision-making are not realized in inner speech since there are too many and they are often not conscious. Second, it contradicts the experience that we often have difficulty in finding the appropriate way to express our thoughts: In these cases we normally know what we think, but the words to express it do not come to mind. This phenomenon would be mysterious if all thought were inner speech. Third, there is empirical evidence that object categorization is not language dependent (Malt, Sloman and Gennari, 2003). Fourth, there seem to be pathological cases in which inner speech is completely disrupted yet thinking seems relatively unaffected (Levine, Calvanio and Popovics, 1982). Fifth, there is a long list of evidence of animal competences that is best explained by presupposing concept possession independent from linguistic competence. Concepts as constituents of thoughts are independent from speech competence (Allen, 1999; Glock, 2000). All this points to the conclusion that thoughts are not wholly dependent on language (inner speech). However, this is not to say that thoughts are not often accompanied by inner speech.[2] We should,

[2] Indeed, it may seem reasonable to take inner speech as one form of thought. But if it is true that speech is only a way to express thoughts (as we believe), then every inner speech act must be preceded (and caused) by a thought. Thus, the experience of a thought as inner speech does not allow for an identification of the two.

nonetheless, distinguish between thoughts, on the one hand, and speech as an expression of them, on the other hand. Because of the frequent co-occurrence of thought and inner speech, some cases of auditory hallucination may be reported as thought insertion (cf. Frith, 1992, p. 71). We propose to classify these cases as delusion of inner speech insertion rather than thought insertion, because we think that true cases of thought insertion—which exist—require a different explanation and hence constitute a class of its own (we will come back to this point in section 3.2).

All in all, Frith's model seems to be a very good model to explain various abnormalities in action control and action awareness. The comparators assumed also give rise to the feeling of agency, that is, the 'labelling' of an action as mine. Nevertheless, the comparator model neither accounts for the basic self-world distinction nor for thought insertion. In both cases we offer arguments to establish the limits of the comparator model, and outline positive accounts explaining the relevant phenomena.

2. Self-World Distinction

One of the fundamental conditions for self-consciousness is the ability to make a distinction between the self and the rest, that is, the world. This self-world distinction is not just given, since every sensation carries information about both the sensing system and the world. Consider proprioceptive afference, for example, while lifting an object. The representational content of this afference comprises information about the movement and position of the limb, but also information about the weight (and position) of the object. In order to distinguish between the weight of the object and the weight of the limb, the force needed to lift the hand alone has to be subtracted from the actual force needed for the lifting movement. Similarly, in order to distinguish whether the world around me has moved or my eyes have moved, the visual afference has to be processed further, since the representational content of the visual input is the same for both cases. The self-world distinction therefore requires some kind of division of input sensation into self-related and world-related information. Because this ability is so fundamental to all aspects of self-awareness or self-consciousness, it can be said to create a core self.

Legrand (2006) has argued that the basic self-world distinction is made by the comparator, such that a pre-reflexive bodily self is constituted by the action monitoring system (control loop; see Figure 2). Whenever the efference copy and the afference match, the afference is taken to be information about the system itself, whereas in the cases of mismatches it is taken to be information about the world. The basic idea is that the bodily self comes into being as an effect of distinguishing the self and the world. The comparator does not label the source of perceptual information. Rather, the self is constituted by the integration of action and reafference performed by the action monitoring system (cf. Legrand, 2006, section 10). However, we will argue that the action monitoring system based on

the comparator-model presupposes a self-world distinction and hence cannot be constitutive for it.

2.1 The Presuppositions of the Comparator-Model

The comparator computes the difference between an efference copy and an afference. It is not just implementing a conditional of the form: if there is a motor command, then take the afference to contain information about me. This means it is not just comparing the presence of an efference with the presence of an afference. It rather compares the representational content of the efference copy and the afference. Otherwise, its monitoring function for specific movements could not be fulfilled, since it does not only matter *that* a movement is made, but it also matters *what* specific movement it is.

In programming languages (e.g. C, Lisp), there are three different equalities of variables. The first one requires that the value of two variables is stored in the same physical place, the second holds if the values of the two variables have the same code, and the third holds if the values are of the same content even if differently coded (e.g. 'a', coded as 61 in hexadecimal code, and 'A', coded as 41, would be such two values). The first two kinds of equalities could be easily computed by a comparator, simply by comparing (i.e. subtracting) the two memory addresses or the two codes. However, the third kind of equality requires that the system 'knows' what counts as the same content. This means that there has to be some table where all codes with the same content are grouped together. In more general terms, if contents are computationally compared, their code will be compared. If, as for the third equality, the comparison should yield a match although the codes are different, the codes have to be 'interpreted', i.e. some table that groups the according codes has to be used. If it is true that neurons do nothing else than computation, this applies to the brain as well, since a neuron can 'distinguish' codes (activation potentials) but not contents.

For the comparator model this means that it has to presuppose that efference and reafference are in the same code.[3] Indeed, there is strong evidence for an intermodal (or amodal) 'common code' of motor signals and perception, which is not only very plausible for proprioception, but also supported by empirical evidence for other modalities (Hommel, Müsseler, Aschersleben and Prinz, 2001; Gallese and Metzinger, 2003; Gallese, 2003). We are not going to argue against this. Quite on the contrary, we will give an account of how such a common coding becomes possible. What we doubt is that such common coding can be assumed to be innate. Consider again the case of arm lifting: In order to be computable for the comparator, the efference copy has to be in the same code as the reafference, i.e. the proprioceptive information of the position of the limb

[3] It is the equivalent to say that the codes are the same and the operation is subtraction, and to say that one code is the inverse of the other and the operation is juxtaposition (addition).

(after the movement). Since the position of the limb after the movement is dependent on the weight of the limb, this weight has to be (implicitly) represented in the efference copy as well. In our view, it is not plausible to assume that the weight of one's own limbs is 'known a priori', since it changes dramatically during development. It has to be learned and recalibrated reflecting growth.

However, it might be argued that the case of lifting objects (and feeling their weight) is rather complicated and yet not crucial for a self-world distinction. The claim that the motor command to the eyes is innately in the same code as the movement registered by the visual system seems much less implausible. It seems that intuitions concerning innateness in this case go in different directions, and we cannot present a conclusive argument to rule out that this kind of common code is innate. What we do offer is an alternative approach that explains how the common code can develop on the basis of other mechanisms, thereby giving an explanation of how comparators can evolve. We take the fact that our account is presupposing less to be innate to be the best argument for it.

Before turning to our alternative account, let us subsume the line of argument given. The comparator presupposes—to be computationally adequate—a common coding (or a table grouping the different codes containing the same content as the basis of a common coding). Common coding (or the table) contains implicitly a self-world distinction, since it involves 'knowledge' of the effects of *my* movements (on my afference). To say that the coding of the information from the visual flow that the world turned 30° to the left is the same as the coding of the motor command to move the eyes 30° to the right, is to say that the effect of *my* movement on my afference (visual flow) is 'known'.[4]

2.2 Perceptual Self-Acquaintance

Once common coding or the relevant table grouping the different codes with same content is established, a comparator can easily be implemented. The table itself involves essentially a basic self-world distinction. The goal of this section will hence be to show how such a table (or common coding) can be established, i.e. to present an account of how a basic self-world distinction can be drawn.[5]

Together with Jan Restat we developed an alternative account of the self-world distinction (Vosgerau, Restat and Newen, 2005). In order to distinguish the self-related parts of the information from the world-related parts, a system needs to have some further information about what perceptual information counts as self-related. A naturalistic explanation of the self-world distinction has therefore to present the source of this information in the system. Since both efferences and

[4] Accordingly, Hommel, Müsseler, Aschersleben and Prinz (2001) admit that their theory 'is meant to provide a framework for understanding linkages between (late) perception and (early) action, or action planning' (p. 849), which is likely to be learned rather than innate.

[5] For a discussion of non-basic forms of self-world distinction see Newen and Vogeley (2003).

afferences are nothing but action potentials of neurons, the self-information cannot be found in the difference between them.

As shown by O'Regan and Noë (2001), systematic contingencies are the source of information for perceptual categories. According to their theory, perceiving an object means to 'know' the systematic covariation between motor actions and change in sensational input. The visual input of a red surface (on the retina), for example, changes when we move around because of changes in illumination. However, we do not perceive these differences but rather a uniform colour red. From the moment we have categorized the visual input as red we 'know' how it will change and therefore are able to nullify changes in the input (with the help of a comparator). The result is a uniform perception of the colour red. In other words: Once we have learned the systematic contingencies between motor action and changes in sensory input the relevant table (or common coding) can be established. The systematic contingencies, however, occur in the system independently of whether they are processed or not.

The detection of systematic contingencies does not involve any kind of comparison. Regardless of the representational content or even the coding of an activation, it can be detected to occur systematically together with some other activation. For detection, the implementation of a simple conditional suffices: if there is a certain activation here, there is a certain activation there. If a system manages to detect these systematic contingencies, it is able to build a table that groups the 'activation there' by what 'activation here' occurred. If the systematic covariation of a certain afference (e.g. the proprioceptive information of my limb's movement) with certain efferences (the according motor commands) is detected, this 'knowledge' can be used to group the efferences to one class corresponding to a certain movement. This table (the 'knowledge' of the systematic contingencies) can then serve as the basis for a common coding on a higher level.

In order to establish a basic self-world distinction, it is therefore necessary to detect systematic contingencies between motor commands and afferences. Such systematic contingencies are not that easy to find, since a passive movement causes pretty much the same proprioception as an active movement. However, this is not quite true: Among our sense organs there are tendon receptors within the tendons of each muscle. As opposed to other muscle receptors, tendon receptors fire only when the muscle is contracted (cf. Restat, 1999). When a muscle is passively moved, the muscle receptors will transfer information about the movement. Since in a (completely) passive movement no contraction of the muscle is involved, the tendon receptors will not respond.[6] Since muscle contraction always involves an efference of the system (action), there is a systematic covariation between efferences and tendon receptor afferences. This covariation can be detected by the system and

[6] The tendon receptors fire constantly; they respond to stretching with a higher frequency. Of course, tendons are also stretched when the muscle is passively stretched heavily. However, such heavy passive stretching is not likely to occur in normal child development and can therefore be disregarded.

carries information about who initiated the movement (the system itself or the 'world').

On the basis of this circle of motor commands and tendon receptor responses, other systematic contingencies arise. There are changes in the sensation that are systematically accompanied by this circle and others that are not, the first being self-caused (re-)afferences and the second being world-caused afferences. Sensation can hence be divided into two classes: the class that is caused by the system itself and the class that is caused by the world. This distinction is the basis for what we called a table: The efferences can now be grouped with their caused reafferences. This table is the basis of the establishment of a common coding of perception and action.

The presuppositions of this explanation of the self-world distinctions are only two: First, the system must start to move somehow; this means it has to show spontaneous (non-intended, non-goal-directed) movements. Children are shown to perform such spontaneous movements long before birth. Second, the system must have the ability to detect and store systematic covariation, thereby creating a table grouping the efferences with the appropriate reafferences. This ability is a system-inherent feature of neuronal networks that simply register systematic contingencies.

Our explanation of the self-world distinction through perceptual self-acquaintance thus relies on unproblematic presuppositions. No intentions and no prior self-world distinction is involved. As soon as a system has established a self-world distinction in this very basic sense, it can develop a comparator and learn to control its movements and to interpret its sensations in terms of self-caused and externally caused afferences. In this way, our account is an explanation of how the common coding of perception and action can evolve within a single system. It does not rely on the implicit innateness of a core self. Therefore, it is a supplement to the common coding view and the comparator model rather than an alternative approach.

3. The Ownership of Thoughts

One of the classical problems of self-consciousness is the feeling of ownership of one's own states. As already mentioned, the comparator-model provides a good explanation for the ownership of action. However, whether this model could be applied to thoughts as well is currently discussed in the philosophical literature. Building on the basis of Frith's theory of schizophrenia, Campbell (2004, 1999) developed a philosophical account of the ownership of thoughts. The basic idea is to define thoughts as motor processes and hence to propose the same comparator mechanism for thoughts and action.

However, there are several difficulties with this view as pointed out by Gallagher (2004*b*). First, an intention to think must be presupposed, which is hard to characterize and contradicts the phenomenology of unbidden thoughts (thoughts

that one does not intend to think). Second, even if the comparator-model could explain the feeling of ownership of thoughts, it cannot explain the misattribution of thoughts to others. Third, there are problems with the data itself: Positive syndromes of schizophrenia are typically episodic, i.e. they do not occur 24 hours a day. Moreover, thought insertion is often limited to thoughts with a certain content. It is hence improbable that thought insertion can be explained by the dysfunction of one module.[7]

Moreover, it is not clear what function the 'thought-comparator' should have besides generating the feeling of ownership. If there is no other cognitive function it is hard to explain why it developed at all. Campbell proposes that '[…] part of the role of the comparator here is to help to keep your thinking on track' (Campbell, 2004, p. 7). Certainly, in some cases, our thinking is directed towards a goal. Consider, for example, a child seeing a piece of chocolate on the table. Presumably, she will entertain a thought that she wants to have that piece. She will then continue with a thinking procedure aimed at figuring out the best way to achieve her goal. Thereby, she will keep her thoughts 'on track', as described by Campbell. However, it is clear that there are two thoughts that have to be compared by the comparator. First, the thought that she wants to have the chocolate, and second, the thought that she will get it by doing this and that. This means, that there must be a thought (as the intention to trigger the process) that can than be compared with the outcome. But how does the feeling of ownership arise for this first thought? It seems that there must be a further thought to get the comparator working in order to create the feeling of ownership. This further thought presupposes another one, and so on.

In our view, these considerations lead to the most important critique. It is the threat of an infinite regress. As Gallagher (2004*b*) remarks, this problem is quite obvious when the intention to think is itself conscious: It is hard to see how a conscious intention to think some thought could not be classified as a thought itself (just like other conscious intentions, e.g. the intention to lift my arm). But if it would itself be a thought, then a further intention (thought) would be needed to produce this one, and so on. However, there may be a deeper problem. The content of the thought *p* has to be present in the intention, for otherwise the comparator could not match the intention with the actual thought in the stream of consciousness (see also section 2.1). Campbell's view thus leads to the conjecture that there are two different mental representations with the same content *p*: The intention, on the one hand, and the thought, on the other. Even if we assume—as Campbell (1999) does—that the intention to think is produced by a sub-personal cognitive module and therefore not available to consciousness, still the status of this intention is unclear. It seems that it should be characterized as a thought itself, that is merely brought to consciousness by some later mechanism (the motor

[7] Unless there are different comparators for different thought contents, which would be highly implausible.

process in Campbell's picture), since the only difference between the two representations seems to be that one is conscious and the other is not. However, if the intention is an (unconscious) thought, and all thoughts are motor processes, then there must be another intention (unconscious thought) to produce this intention. This still leads to an infinite regress.

If the unconscious intentions are not classified as thoughts, another problem arises. For every thought there must be an unconscious intention with the same content. Hence, in thinking processes, all work has to be done at an unconscious level. Thinking itself would then be best described as the movement from one unconscious intention to another unconscious intention. The fact that these unconscious intentions cause some thoughts in the stream of consciousness turns out to be a mere epiphenomenon, which contradicts the normal phenomenology of thinking. More important, however, is the fact that in this picture the original function of the comparator—namely keeping thoughts on track—could not be fulfilled. For this function, it is necessary that one thought can be matched with some previous thought in order to test if there is a track. However, first the comparator does not compare different thoughts (since this leads to an infinite regress; see above) but rather unconscious intentions and thoughts. The same applies to unconscious intentions: They are not compared and therefore they cannot be kept on track either. Second, the content of the intention and the content of the thought have to be the same if they should match, which means that the comparator could only keep the same thought on track.

The comparator model for thoughts thus leads to a dilemma: Either we characterize intentions as thoughts and are confronted with an infinite regress, or we characterize them as some unconscious representations, which leads to the conclusion that the function of the comparator (keeping track of thoughts) cannot be fulfilled. Moreover, the second horn of the dilemma also leads to the view that (conscious) thinking is a mere epiphenomenon, and here the function of the feeling of ownership becomes unclear as well.

3.1 The Difference between Thoughts and Motor Processes

Faced with these problems, the question whether thoughts can be characterized as motor processes imposes itself. We will first take a closer look at how motor processes are described by Frith (1992, 2000). The second step is the discussion of the transferability of this picture to thoughts.

The monitoring system requires—according to Frith (see Figure 2)—three matches to be made. One is the comparison between the desired and the actual state (feedback loop), another is made between the desired and the predicted state (feedforward loop), and a third occurs between the predicted and the actual state (control loop). The first comparison carries the information of success, thereby providing the feeling of ownership ('It was me, who did that' as opposed to 'Some external force was involved'). The second comparison monitors the adequacy of the movement before it is executed, thereby providing the feeling of agency,

i.e. the feeling of producing and controlling the action. The third comparison allows for fast on-line correction of the movement independent of visual feedback.

The feedforward loop involves the comparison between (the representations of) the desired and the predicted state. In order to be comparable, the two representations must have the same content (see section 2.1). Therefore, since the content of the desired state is the result of the movement, the content of the predicted state is as well. On the other hand, however, the predicted state is directly compared with the actual state in the control loop. The control loop requires a continuous comparison during the whole movement since otherwise the fast on-line corrections could not be explained. Hence, the content of the predicted state representation is twofold: It involves a static representation of the goal state as well as a dynamic representation of the whole movement. Because of its latter content, it is assumed to be an instance of imagery of movements as well (Jeannerod, 1999). Imagery of movements is then exactly the same as performing a movement except for the blocking of the motor commands. Since the feedforward loop is working in imagery, mental training of movements is possible.[8]

If thoughts were motor processes, the picture would be like this: Some intention to think *p* triggers a desired thought *p* that is the input of a thought generator. The thought generator specifies the thought, sending specific thought commands to the stream of consciousness and thereby producing a predicted thought *p*. What we call the thought *p* is only that part occurring in the stream of consciousness (parallel to movements). Besides the critique already mentioned, this picture leads to even more problems. The advantage is that the dissociation between ownership and agency of thoughts can be explained. However, the question what an intention to think *p* should be can be sharpened now. As in motor processes, the desired thought *p* should be a representation of the end product of the thinking process, whereas the predicted thought *p* should represent the whole process of thinking. The actual thought *p* occurring in the stream of consciousness would be the actual process of thinking. The main problem here is that thoughts are intuitively not describable as processes (though thinking is). The thought 'Trees are green' is not a process but rather the end product of a thinking process. Moreover, the unconscious 'background beliefs, desires, and interests, together with current external stimuli' (Campbell, 1999, p. 617) that cause our occurrent thought are involved in the process of thinking. If it were the process to be found in the stream of consciousness, they would be conscious as well. Therefore, what is meant by the term occurrent thought is rather what appears to be the desired thought in this picture. Moreover, imagining thoughts without thinking them should be possible, which is clearly implausible.

---

[8] This picture implies that movement imagery should give us the feeling of agency but not the feeling of ownership, whatever this means. However, this problem is not within the scope of this paper.

Given these considerations, it seems that thoughts cannot be characterized as motor processes.[9] If the plausible theory of motor processes by Frith is applied to thinking, various highly problematic and implausible, if not inconsistent, implications have to be drawn. We will therefore turn to an alternative picture of thoughts and their relation to motor processes that better captures our intuitions. We then sketch an analysis of how the phenomena of thought insertion can be explained within the new framework.

3.2 A New Framework: Thoughts as Intentions

Thinking is doubtless a process. However, thoughts are rather the products of thinking processes. Thinking can then be described as proceeding from one thought to the next. There seem to be various mechanisms to be involved in thinking processes such as association, deduction, induction, analogical reasoning, and so on. However, we will concentrate on thoughts and their relationship to motor processes, leaving thinking largely apart.

Certainly, thoughts can play an explanatory role for behaviour. My thought 'There is a restaurant which provides good food' will be part of the explanation of my walking there. The intention to walk to the restaurant is therefore (partly) dependent on this specific thought. We take this intention to be the one that triggers the desired state and herewith the whole motor process. What is then added to the thought to become an intention? Where exactly does the difference lie between thoughts and intentions? Of course, the thought 'There is a restaurant' is not sufficient to trigger a certain kind of behaviour. There must be also the feeling of hunger or appetite, the knowledge of having enough money to pay, the knowledge of having the necessary time, and so forth. All these thoughts may actually occur in the thinking process actually entertained. Nevertheless, there will be one thought as the end product of this thinking process, namely the thought expressed by 'I will go to the restaurant'. If this thought does actually trigger the motor process, it becomes the necessary intention.[10]

We can therefore describe intentions as thoughts that trigger a certain motor process. In other words, the functional role of the thought (triggering motor processes) qualifies it as an intention.[11] There might be other functional roles of thoughts, for example being the premise for another thinking process. Moreover, there might be other sources of motor processes, such as in reflexes or automated actions. However, in this picture, thoughts are essentially different from motor actions (see Figure 3). They can trigger motor processes. If they do, they thereby qualify as intentions.

---

[9] Of course, utterances are motor processes. However, we sharply distinguish them from thoughts here (see also section 1); we will come back to their relationship in section 3.2.

[10] The thought expressed by 'I will go to the restaurant' is the triggering cause on the basis of several background conditions that are the structuring causes (cf. Dretske, 1988).

[11] We can therefore speak of the propositional attitude of intention (will) that I have towards the thought (proposition) 'I go to the restaurant'.

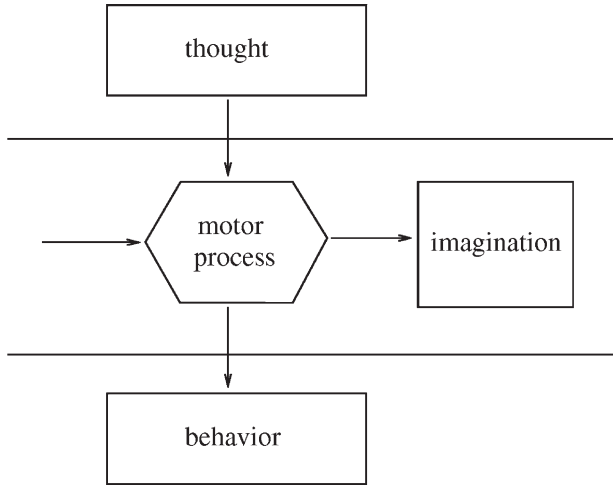

Figure 3 *Thoughts as intentions*

This picture is consistent with the view that intentions do not necessarily lead to overt behaviour. The specified movement commands can be blocked such that the intention results in imagination (of movements). It is likely that we use this possibility of action simulation quite frequently. The feeling that thoughts and imagined actions are not clearly distinguishable may stem from the fact that thoughts are very often accompanied by imagined actions. Nevertheless, these two mental entities must not be confused.[12] Since imagined actions are the predicted states in Frith's model, they have the same content as the afferences (i.e. the perceptual input). Hence, imagined actions are like percepts, unlike thoughts. However, just as a perception may give rise to the grasp of a thought (seeing a tree may lead to the thought 'There is a tree'), imagination can be a source of thoughts. In this way, imagination, i.e. the entertaining of a motor process without overt output, can be part of thinking processes.[13] Nevertheless, imagined actions are not thoughts themselves.[14]

---

[12] Empirical research shows that visual imagination can even impede reasoning processes (Knauff and Johnson-Laird, 2002).

[13] This is not to say that imagined actions have to be part of thinking processes. Deliberating a thought, for example, can be fully performed on the level of thoughts without imagination.

[14] Whereas the primary function of imagination is the anticipation of action, thoughts are primarily intentions, i.e. reasons for action. The difference is especially clear if a cognitive system has self-knowledge. Then, critical reasoning becomes possible, which is a process of considering reasons as reasons and thereby evaluating and selecting intentions. Since imagination can give rise to thoughts, they may be a trigger for critical reasoning as well; however, imagined actions are not themselves deliberated, since they just represent a possible outcome of an action but not a reason for it. The close relation between self-knowledge and critical reasoning is widely agreed on, although its direction is discussed between Burge (1996) and Peacocke (1996).

What about the phenomenon of thought insertion? We will start with a discussion of delusion of control. The feeling that movements are made by me but not controlled by me can be explained by the comparator-model (see section 1). However, this explanation does not comprise the attribution of the control to another person. As Davies, Coltheart, Langdon and Breen (2001) show, for a multitude of psychiatric phenomena there are cases of delusions involving misattribution and non-delusional cases without misattribution. There are patients who say that they do not have the feeling of control and that this is due to a lesion (rather than some alien force). Therefore, the breakdown of comparator mechanisms can explain the loss of the feeling of agency but not the misattribution. For this reason, Davies, Coltheart, Langdon and Breen (2001) conclude that a two-factor account of delusions is inevitable. The first factor has to explain a certain kind of a strange feeling (e.g. actions without the feeling of agency), whereas the second factor has to explain the misattribution of actions and thoughts to external forces.

There is evidence for a process of rationalization in folk psychology and scientific psychology. Whatever actions we perform, we afterwards are ready to give a (more or less) plausible explanation for our actions. This process of rationalization is well supported by psychologists. The famous theory of dissonance (Festinger, 1957, 1964) explains contradictory statements about subjects' behaviour with rationalization processes. If an action does not lead to the result previously rated desirable by the subject, subjects often start to explain why the desired result is not that good and why the actual outcome was the 'real' intended result, thereby contradicting their previous statements. The theory explains this fact by a rationalization of the action. Rationalizing comprises three strategies: adaptation of beliefs, attribution to external sources, and repression. If there is a dissonance, i.e. a conflict between beliefs and perceived states, then the rationalizing process starts in order to remove the conflict. This can be done either by changing the beliefs, by attributing the cause of the state to some external source, or by repressing a belief. If, for example, I have the belief that I put the key on the table but I cannot find it there, I will either change my mind and believe that I put it somewhere else, or attribute the cause of this situation to somebody else by believing that someone has taken the key, or by trying to forget about the silly key and repress the whole conflict. The fact that our post hoc explanation often does not fit the earlier statements, makes clear that this mechanism is not based on reliable introspection but is rather an ad hoc way of theorizing about oneself.

If a patient has the feeling of performing some action without having the feeling of agency (due to a comparator breakdown), he will rationalize this phenomenon. The third strategy of repression will not work when this strange feeling occurs over and over again. The first strategy is hard to adopt, since both contradicting beliefs ('I moved' and 'I didn't control this movement') are both grounded in very reliable evidence: the first arises on the ground of perception (extero- as well as proprioception); the second is based on a strong introspective feeling. Changing

the beliefs would therefore mean either believing in having perceptual hallucinations or believing that the subjects' feelings are wrong, both of which are very costly. Hence, the normal way of rationalizing leads to a misattribution of the cause of the action. The second factor for delusion of control is hence a normal rationalization process, a process that is well described for healthy subjects and different problems.[15]

The case of thought insertion is more complicated, not only because there is no mechanism providing either a feeling of agency or of ownership in our picture. It is also the astonishing lack of data (see Frith, 2004) that makes a neat description of the phenomenon difficult. It is not at all clear what feelings patients try to express by stating that they have thoughts that are inserted into their mind. In a way, thoughts are inserted into your mind when you read this article. However, this form of 'thought insertion' does not produce a strange feeling (we hope). As Gallagher remarks, '[t]his frequently happens in academic situations as I listen to a lecture and thus allow someone else to guide my thought' (Gallagher, 2004*a*, p. 91). Campbell (1999, p. 620) discusses whether 'inserted thoughts' make sense at all. Indeed, there seems to be a sense in which thoughts are produced by me, and there seems to be a knowledge about that. However, the question is whether there is a *feeling* of agency of one's own thought. We doubt that. It seems that there is no phenomenal difference between the thought that 2 + 2 equals 4 grasped in a math class or grasped in a supermarket while calculating the prices. Indeed, it is even hard to say that this thought is mine in the sense that the idea of quanta is Einstein's. Equally, there is no phenomenal difference between an unbidden thought that just occurs to me and a 'purposeful' thought with the same content. So-called inserted thoughts occuring in healthy subjects are also not accompanied with a special feeling of the source. We therefore think that any bottom-up account (as proposed by Gallagher, 2004*a*) that explains misattribution by the feeling that somebody else causes the thoughts, fails. It fails not because it cannot account for the data in schizophrenia, but because it cannot explain the lack of the feeling in healthy subjects. It seems, instead, that 'ownership of thoughts' is some sort of background knowledge.

We therefore take it that the 'second factor' of thought insertion (creating the misattribution) is a normal and healthy rationalization, just as in the case of delusion of control. In fact, the so-called feeling of ownership of thoughts in healthy subjects is some kind of background knowledge that is constituted by the same rationalization processes. Therefore, ownership of thoughts is not a basic feeling; it is itself a product of forming a theory about one's own mental life. In the case of thought insertion, only the 'first factor' is pathologically impaired, such that it gives rise to some kind of strange feeling about one's own thoughts (thereby creating the dissonance to be rationalized).

---

[15] We have to admit that we cannot explain the fact that some patients are open to the better explanation of the psychiatrists and some are not (i.e. have delusions). We can only speculate that the confidence in the subjects' feelings differs across persons.

We suggest that there are at least three possible sources of the feeling of so-called inserted thoughts:[16] The feeling may be based on impairment of those mechanisms that are usually combined in a standard way with the process of thought production. The relevant mechanisms are those (i) of speaking, (ii) of imagination, or (iii) of emotional evaluation. The first one involves the motor processes of speaking. There are several experiments showing that some patients claiming to have so-called inserted thoughts speak out those thoughts unconsciously (see Frith, 1992, p. 71f).[17] It is likely that such patients have problems with the motor control of speaking in the sense of Frith—while the act of speaking is combined with thought production—i.e. this case can be explained on the basis of the comparator-model with an impairment of the predicted state involved in speaking, resulting in a mismatch at the feedforward comparator (see Figure 2). These patients are better described as having auditory hallucinations. The second source involves imagination that is combined with thought production, especially imagining speaking (inner speech). Here, the motor process causes imagination of an utterance but not a speech act. This case is closely related to the first case. The only difference seems to be, that in the latter no subvocal speech is involved. These cases can also be described as auditory hallucinations (without subvocal speech). There is evidence that in both types of hallucinations brain areas are involved that are associated with speech generation (the 'specification of movement' module in Figure 2) (Stephane, Barton and Boutros, 2001). It is likely that the feeling of inserted speech (inner speech hallucination) is expressed by some patients by declaring that thoughts have been inserted into their mind. However, the mechanisms involved in such cases are different form the mechanisms involved in proper thought insertion (see next paragraph), and hence these cases should be classified as a different phenomenon (e.g. verbal hallucination).

Nevertheless, a third category must be proposed, since 'many people who experience voices are not having auditory hallucinations. They do not mistake their awareness of inner speech for auditory perception of somebody else's speech, nor do they even have the impression that they are hearing another speak. Thus, verbal hallucinations cannot be regarded, in general, as an audition-like experience' (Stephens and Graham, 2000, p. 103). So we need a perception-independent account that 'explains how voices can be experienced as alien without being experienced as auditory' (Stephens and Graham, 2000, p. 103). This case is then the only case that really deserves the name thought insertion.[18] We speculate that

---

[16] This is not to say that we take all of these cases to be thought insertion proper. Indeed, only the third case will qualify as such. However, because of the close connection of the other two to thought production, the first two may be (mis-)reported as cases of thought insertion.

[17] This phenomenon turns out to be very rare; it can be hence maybe viewed as marginal, so that its explanation does not contribute much to the explanation of thought insertion.

[18] Contra Stephens and Graham (2000), we think that in these cases we should not speak of verbal hallucination but rather of thought insertion, since inner speech leads to an 'auditory-like experience'.

in these cases an impairment of the emotional evaluation of thoughts is involved. The claim that we have an emotional encoding in such cases is based on the fact that the delusion of alien thoughts is a stable fact that cannot be changed by information. Since the perceptual system is not involved in the third source, the best candidate being involved is the emotional system. The impairment of the emotional system can result in strong emotional aversion to specific contents. The aversion itself can stem from education, traumatic episodes, or psychiatric disorders.[19] Its overwhelming strength stems from an impaired emotional system that merely differentiates between very good and very bad but nothing in between.[20]

The three different mechanisms all lead to dissonance situations in which two beliefs conflict with each other. In the cases (i) and (ii), patients hear voices that conflict with their beliefs about their environment ('There is nobody speaking'). As pointed out for delusion of control, the normal rationalization processes lead to a misattribution of the source of these verbal experiences. In case (iii), where patients show extreme emotional evaluation, a thought with a specific content will evoke a strong aversion against it. The patient is faced with a dissonance between his occurrent evil thought and his strong belief that he is not evil and not ready to think such evil thoughts (as it is for the case reported in Frith, 1992, in which a woman believes that the thought 'Kill God' is inserted into her mind). Here, the rationalizing process leads to a misattribution in order to 'keep the self clean'. Indeed, thought insertion is very often confined to specific contents.[21] In all cases, the misattribution of the verbal experience or thought through the rationalizing process becomes so prominent that after a certain period patients report a *feeling* of thought insertion and are not ready to give up their explanation in favour of the scientific one.

Our account of thought insertion is a two-factor account, which is 'hybrid' in the words of Gallagher (2004*a*). There are three different first factors that give rise to dissonance situations, only one of which is involved in the case we call thought insertion. So far, it is a bottom-up account, and we agree with Gallagher that the abnormality of schizophrenia is found in the phenomenal level. The second factor is the rationalization process that is also at work in healthy subjects. This part of the account is a top-down account. In our picture, thought insertion that is not secondary to auditory or verbal hallucination involves an impairment of the emotional system and the normal rationalization processes. However, there are cases in which the thoughts that are reported as inserted have a rather banal content ('good job', 'OK', etc.). A large number of these cases could be cases of what we called verbal hallucination, and not cases of proper thought insertion. We cannot exclude the possibility of proper thought insertion of banal contents, yet it is implausible that such contents evoke a strong emotional reaction. For those cases,

---

[19] This claim is supported by the fact that the treatment of schizophrenic patients includes conversation therapy with the aim of normalising the emotional response.

[20] For some case descriptions, see e.g. Erichsen, 1973.

[21] The authors are not aware of a single case of general thought insertion.

there may be a further source of dissonance for thoughts that we have not described. One possibility is that this source is to be found in an impairment of the thinking processes, which are not discussed here. However, we have presented a framework for the explanation of thought insertion that can deal with many cases. Some cases may not fit into our account, but for these cases the framework is easily and straightforwardly extendable.

4. Conclusion

The comparator-model (Frith, 1992; Frith, Blakemore and Wolpert, 2000) provides a good explanation of motor processes. In particular, it can explain certain pathological phenomena such as delusion of motor control. However, we have shown that it can neither be extended to explain a basic self-world-distinction nor to explain the phenomena of thought insertion in schizophrenia.

A basic self-world distinction can be made on the ground of systematic contingencies between motor commands and tendon receptor activation. Whenever an organism actively moves, this efference-afference circle occurs. Hence, changes in the perceptual flow can be divided into two classes: those accompanied with this circle (self-caused changes) and those that are not (world-caused changes). In this way, a basic self-world distinction can evolve which builds the basis for a common coding of action and perception and thereby for comparators.

Regarding thought insertion, with its astonishing lack of data, we can only present a vague idea that has yet to be further specified and empirically tested. Nevertheless, the core of our argumentation is that thoughts cannot be characterized as motor processes. Rather, thoughts differ essentially from motor processes and imagination, because they can be the triggering cause (intention) for the latter. Moreover, there are other functional roles of thoughts, for example being the premise for a thinking process. All in all, a lot remains to be said in order to characterize thoughts precisely. However, it is clear that they cannot be characterized as motor processes.

The explanation of delusions requires a two-factor account. One factor has to explain the strange experiences patients claim to have, while the other factor has to explain the misattribution of actions and thoughts. We suggest the second factor is a normal rationalization process. Healthy humans rationalize their own behaviour as well as their own mental states, i.e. they try to find a good explanation for it. If they cannot find such an explanation, they tend to repress what happened or to attribute it to some external force. In this way, a coherent picture of the self can be created and thoughts can be classified as being the subjects' thoughts. However, if two contradictory beliefs—produced by the strange experience—occur over and over again, the source for one of the beliefs is (mis-)attributed to an external source.

We presented three possible sources of the 'strange feelings' that are rationalized. First, the rare case of speaking subvocally could be reported as thought insertion although it is rather a case of alien control phenomenon. Second, imagination (inner speech) can be experienced as not belonging to oneself in case of a comparator breakdown. These cases are not cases of thought insertion in the strict sense, although they might be often reported as such by the patients. The third case is the only case of thought insertion in the strict sense, where the emotional evaluation of thoughts grades them as evil, which contradicts the patients' belief not to be evil. This list is probably not exhaustive, and there might well be many more sources of such strange feelings. However, the first step to finding this out is to present considerable criteria for classifying the diverse cases of so-called thought insertion. We hope that this paper can serve as a first step towards a framework that allows for such a thorough classification.

Philosophisches Seminar
Universität Tübingen

References

Allen, C. 1999: Animal concepts revisited: the use of self-monitoring as an empirical approach. *Erkenntnis*, 51, 33–40.

Burge, T. 1996: Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society*, 46, 91–116.

Campbell, J. 1999: Schizophrenia, the space of reasons, and thinking as a motor process. *The Monist*, 82, 609–625.

Campbell, J. 2004: The first person, embodiment, and the certainty that one exists. *The Monist*, 87, 475–488.

Davies, M., Coltheart, M., Langdon, R. and Breen, N. 2001: Monothematic delusions: towards a two-factor account. *Philosophy, Psychiatry, and Psychology*, 8, 133–158.

Dretske, F. 1988: *Explaining Behavior*. Cambridge, MA: MIT Press.

Erichsen, F.. 1973: Die Bedeutung von Werthaltungen für die Schizophrenie [The significance of valences in schizophrenia]. *Psychiatria Clinica*, 6, 30–52.

Festinger, L. 1957: *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Festinger, L. 1964: *Conflict, Decision, and Dissonance*. Stanford, CA.: Stanford University Press.

Frith, C. D. 1992: *The Cognitive Neuropsychology of Schizophrenia*. Hillsdale, NJ: Erlbaum.

Frith, C. D. 2004: Comments on Shaun Gallagher. *Psychopathology*, 37, 20–22.

Frith, C. D., Blakemore, S.-J. and Wolpert, D.M. 2000: Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London B Biological Sciences*, 355(1404), 1771–1788.

Gallagher, S 2004*a*: Agency, ownership, and alien control in schizophrenia. In D. Zahavi, T. Grünbaum and J. Parnas (eds), *The Structure and Development of Self-Consciousness*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gallagher, S. 2004*b*: Neurocognitive models of schizophrenia: a neurophenomenological critique. *Psychopathology*, 37, 8–19.

Gallese, V. 2003: A neuroscientific grasp of concepts: from control to representation. *Philosophical Transactions of the Royal Society of London B Biological Sciences*, 358(1435), 1231–40.

Gallese, V. and Metzinger, T. 2003: Motor ontology: tthe representational reality of goals, actions and selves. *Philosophical Psychology*, 16, 365–388.

Glock, H. J. 2000: Animals, thoughts, and concepts. *Synthese*, 123, 35–64.

Grush, R. 2004: The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27, 377–442.

Helmholtz, H. 1866: *Handbuch der Physiologischen Optik*. Leipzig: Voss.

Hommel, B., Müsseler, J., Aschersleben, G. and Prinz, W. 2001: The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, 24, 849–937.

Jeannerod, M. 1999: To act or not to act: perspectives on the representation of actions. *The Quarterly Journal of Experimental Psychology*, 52A, 1–29.

Knauff, M. and Johnson-Laird, P. 2002: Visual imagery can impede reasoning. *Memory and Cognition*, 30, 363–371.

Legrand, D. 2006: The bodily self: The sensori-motor roots of pre-reflexive self-consciousness. *Phenomenology and the Cognitive Sciences*, 5, 89–118.

Levine, D., Calvanio, R. and Popovics, A. 1982: Language in the absence of inner speech. *Neuropsychologia*, 20(4), 391–409.

Malt, B. C., Sloman, S. A. and Gennari, S. 2003: Speaking vs. thinking about objects and actions. In D. Gentner and S. Goldin-Meadow (eds), *Language in Mind: Advances in the Study of Language and Thought*. Cambridge, MA: MIT Press.

Newen, A. and Vogeley, K. 2003: Self-representation: Searching for a neural signature of self-consciousness. *Consciousness and Cognition*, 12, 529–543.

O'Regan, J. K. and Noë, A. 2001: What it is like to see: a sensorimotor account of vision and visual consciousness. *Synthese*, 192, 79–103.

Peacocke, C. 1996: Our entitlement to self-knowledge. entitlement, self-knowledge and conceptual redeployment. *Proceedings of the Aristotelian Society*, 46, 117–158.

Restat, J. 1999: *Kognitive Kinästhetik. Die modale Grundlage des amodalen räumlichen Wissens*. Lengerich: Pabst Science Publ.

Sperry, R. W. 1950: Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, 43, 482–489.

Stephane, M., Barton, S. and Boutros, N. 2001: Auditory verbal hallucinations and dysfunction of the neural substrates of speech. *Schizophrenia Research*, 50, 61–78.

Stephens, G. and Graham, G. 2000: *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*. Cambridge, MA: The MIT Press.

von Holst, E. and Mittelstaedt, H. 1950: Das Reafferenzprinzip. *Die Naturwissenschaften*, 20, 464–476.

Vosgerau, G., Restat, J. and Newen, A. 2005: Perceptual self-acquaintance. *Psyche*, 11(1), (abstract: http://psyche.cs.monash.edu.au/ASSC8/ps2004.html, poster: http://www.uni-tuebingen.de/selbstbewusstsein/files/assc8.pdf).