Holk Cruse

Foundations of Feelings

1. Introduction

Internal states like emotions were long regarded as outside the scope of the natural sciences because they are subjective entities and not accessible to measurement. This separation was less strict in 19th century biology, which led to anthropomorphisms later rejected. Behaviorism can be regarded as the most prominent program introduced to counteract such naive approaches. However, the use of introspection was not completely discarded. Application of psychophysical methods provided important results for many decades. In this research, the human subject was used as a measuring instrument, so to speak. This is suitable when, for example, the questions investigated can be formulated to take a yes-no response (e.g. is stimulus A smaller or greater than stimulus B?). It is not clear, however, how these methods can be easily applied to the investigation of emotions like fear, joy or the like. On the other hand, in recent years the impact of emotional states on many if not all types of behavior (e.g. learning) has been increasingly discussed and acknowledged (Cabanac 1992). Furthermore, many psychologists and philosophers argue that emotions are not merely interesting epiphenomena of our life, but form the center of our world. There can be no subjectivity and no sense of the world without our emotions. And even in the field of robotics, robots' lack of ability to experience emotions is seen as a hindrance to constructing really autonomous robots.

Introspection has been the interest and method of the phenomenologists. Although they usually consider themselves to be working in the humanities, and I would not dare to regard them as full-blown natural scientists, they can be seen as laying the foundations of an early state of science in which only qualitative and descriptive methods are available yet. One of these philosophers, H. Schmitz, has spent many years developing a systematics of human emotions, what he calls a phenomenology of bodiliness (corporeality, *Leiblichkeit).* Although Schmitz seems to consider his results inaccessible by the methods of the natural sciences, in the following I attempt to compare his results with knowledge from neurobiological research and the computer modelling of neuronal systems. Section 2 provides a short summary of the coordinate system H. Schmitz developed to systematically describe the human feelings. In

Section 3 a transformation of this coordinate system will be proposed. This new coordinate system is not meant as to replace Schmitz' system but rather to provide a hypothetical bridge to a physiological interpretation. Sections 4 and 5 discuss the question of whether such physiological systems may have a 1st person's view. Section 6 shows how this system may be realized in the form of an artificial neural network.

2. Hermann Schmitz' system of feelings

The following summary of H. Schmitz' findings is necessarily much less detailed than in his original publications (Schmitz 1965, 1966, 1969, for a short version see 1989), but, I hope, contains the essentials. As a first step toward explaining his system, the notation of "bodily islands" *(Leibesinseln)* should be mentioned. To understand these observations, the reader is asked to perform the following two exercises (the first being much easier than the second): when in a relaxed situation, one can look at or touch with one's own fingers the whole body and thereby experience the body as a continuous object. However, in the same relaxed situation, one can close eyes and study the „automatic" feelings of one's body (automatic means here not actively concentrating on special parts of the body, but permitting a more "passive" consciousness). In this situation, only some individual sections of the body are felt. Some of these regions are mechanically stimulated (e.g. the thighs when sitting on a chair), but there are also others where such stimulation does not obviously exist, for example the lips. H. Schmitz calls these regions bodily islands to stress their spatial and discontinuous character. This separation becomes even more obvious in the case of phantom limbs, where usually not the whole limb but isolated parts, for example the hand, are often experienced.

In the following we shall examine other subjective experiences, psychological states or feelings such as joy, sadness, depression (or melancholy, *Schwermut)*, dullness *(Mattigkeit)*, freshness *(Frische)*, fright *(Schreck)*, pain, or dozing. Here we will speak only of those situations in which one is really touched by these feelings (e.g. the sadness arising when a good friend has died, rather than indirect experience such as condoling somebody for the death of a person one didn't know oneself).

These experiences which will be called feelings here can be distinguished into two overlapping domains, the emotional feelings *(Gefühle)*, and the bodily feelings *(leibliche Regungen)*. H. Schmitz' description of such subjective experiences starts with the observation that all these experiences have a spatial character. The bodily feelings lay hold of

parts of the body or sometimes of the whole body, but are always localized within the body. Emotional feelings also have a spatial character, but they seem to extend beyond the limits of one's body. This spatial extension may be described with the metaphor of a diffusing gaseous substance, i.e., there are no defined borders of this space, or better of the "volume" of a sound, or of the feeling one has in a specific climatic situation, e.g. an oppressive climate. In such a climatic situation, the feeling is related to depression. In a fresh atmosphere, one feels happy. Examples of observations leading to these ideas are: Feeling sadness is like being in a low space, joy like in high room (we feel free, or lifted, or as if we were in a high hall). Joy makes all movements easier. Sorrows depress as if they were physical masses. Fright concentrates narrowly in the upper belly, the feeling of dozing is flat and wide.

What is the difference between emotional feelings and bodily feelings? Dullness can be localized, e.g. one speaks of dull legs. This means that, according to the above definition, dullness is a bodily feeling; sadness cannot be localized in this way (there is no sad leg) and therefore belongs to the emotional feelings. Correspondingly, freshness, a bodily feeling, may be localized, in contrast to joy, which therefore is an emotional feeling. As mentioned above, the experience of a bodily feeling is not necessarily identical to feelings produced directly by stimulation of a specific part of the body. For example, a toothache is not restricted to the individual tooth, but rather fills a more extended volume. Feelings can also be elicited by a picture or a movie, which shows that the direct physical effect, for example of a fresh atmosphere, is not the necessary cause for the appearance of the feelings. According to the idea of spatial localization, different bodily feelings can coexist simultaneously. Emotional feelings cover all space, which correspondingly means that they are dominant: one emotional feeling suppresses the others. For example, a fresh person can remain fresh when being together with tired people. However, a happy person appearing in a group of sad people will at least feel uneasy, what H. Schmitz describes as "emotional contrast". (This also shows that, at least in some cases, an emotional space may be shared by different subjects.)

H. Schmitz proposed that all experienced feelings can be described by three dimensions or, in other words, ordered along three coordinates. The first dimension is arranged along a line extending from narrowness *(Enge)* to breadth *(Weite).* For example, in joy breadth is determinant, in sadness it is narrowness.

Usually, as Schmitz observes, there is an antagonistic mutual competition between both tendencies (narrowing, broadening), but the attempt to inhibit the other tendency at the same time serves as an excitation.

What this seemingly contradictory statement means may become clearer by performing another exercise: During inhalation, in particular if one has just arrived at a place of fresh air, one feels that breadth increases. However, during continuation of inhalation, this extension is soon counteracted by an increasing narrowness. During exhalation, one feels neither broadening nor constriction, but merely the resolving of the competition between both tendencies.

Schmitz stated that, during normal conscious living, there is some kind of balance between the two bodily tendencies of broadening and narrowing. Sometimes one tendency dominates the other. Broadening, for example, dominates in inhalation, during pleasurable stretching of the limbs, and joy. Narrowing dominates in fright, fear, concentration, attention, hunger, and anxiety. However, the "bond" between these antagonistic tendencies may loosen. In extreme cases of fright or pain the "bond is broken up", leading to loss of consciousness. Correspondingly, consciousness is lost when broadening becomes absolutely dominant in the transition from dozing to sleeping. Dynamic interaction between the two tendencies plays an important role in Schmitz' description. For the sake of simplicity, here **I** will deal only with a static view. Dynamic effects will not be discussed until Section 7.

The example of inhalation and exhalation also provides an illustration of the second dimension, which **H.** Schmitz calls "bodily direction". During inhalation, in addition to broadening, one also has the feeling of an upward direction. During exhalation, there is no narrowing, but a downward direction. These two dimensions are illustrated in Fig. 2a, using inhalation and exhalation as an example.

The third dimension of H. Schmitz' emotional world is arranged between two tendencies which he calls "epicritical", meaning pointed, bright, sharpening, and "protopathic", meaning obtuse, diffuse, with disappearing borders. This dimension seems to be related to the first one, arranged between breadth and narrowness, but it is independent, because hunger, for example, is narrow but protopathic.

It may be interesting to compare these observations with contentions of Aubè and Senteni (1996), who distinguish in a similar way between two forms of motivational processes that they call needs and emotions. Needs refer to resources that the animal or the human has at his disposal, for example available food or water. Needs are motivational states necessary to control internal states like hunger or thirst. Emotions, like fear, refer to resources whose access depends on others. Emotions, therefore, are used to control social relationships. This may parallel H. Schmitz' distinction between bodily feelings and emotional feelings. The former are restricted to the body of a person. In emotional feelings,

several persons can share a common space, which may lead to mutual excitation or suppression (as in the case of emotional contrast).

3. Comparison with some observations in neurobiology

In the following I shall compare these findings with some observations in neurobiology and later neuroinformatics (see Section 6) and make an attempt to connect these areas.

Recent development indicates that our mental body image is bound to some areas of our cerebral cortex (for example the somatosensory and the motor cortex, e.g. Damasio 1994). It could be shown in different situations that parts of these cortical fields are active in the same way when the subject performs a task as when he or she only imagines performing this task (Kandel et al. 1995). Corresponding results are shown for different sensory parts, for example the visual system. Color vision and color imagination require the same cortical area (Damasio 1994). When this area is destroyed, e.g. by an accident, the patient can no longer imagine what "red" means, for example. If parts of the somatosensory system are destroyed or their function is impaired by indirect effects, then the patient may not be able to imagine the corresponding body part. He cannot accept it as belonging to his body, although he can see it with his eyes (anosognosia, e.g. Sacks 1984).

Because these cortical areas seem to represent the morphological basis of our body image, a simple first hypothesis would be to assume that H. Schmitz' observations are reflected in these cortical areas. Thus, we may hypothetically assume that the experience of the bodily islands corresponds to excitation of the corresponding cortical areas. (It is unclear how the separation of these "islands" is obtained. It might be based on hardwired, e.g. innate connections or might be variable and influenced by activities in other parts of the brain.) Using this hypothetical foundation, the competition between different feelings might be explained by the assumption that the overall excitation of these cortical maps has an upper bound: thus the body islands disappear when a strong, e.g. a painful stimulus arises (Fig. 1). If this stimulus is strong enough, it „consumes" the whole available excitation energy, thus causing other sensations to disappear.

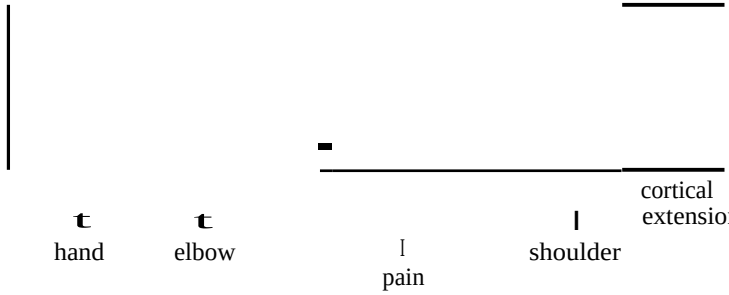

Fig. 1. As an illustration, bodily islands (solid lines) are assumed to occur near the hand, the elbow and the shoulder. The abscissa may be interpreted as a crude representation of the mechanosensory cortex's geometrical extension. The ordinate indicates excitation intensity. If a painful stimulus occurs, this takes over the whole available excitation energy (dashed lines) and the former perception of the bodily islands disappears.

Consideration of the above example of inhalation may lead to the initial assumption that high excitation intensity correlates with the feeling of breadth and low excitation with narrowness. However, this picture does not appear to explain all the observations that Schmitz connects with his notation of broad and narrow. During inhalation, first, breadth is felt to increase, but later the sensation of narrowing becomes dominant. This is indicated in Fig. 2a, using Schmitz' two coordinates breadth / narrowness and bodily direction. However, one also has the feeling that efforts (or energy costs) increase slowly at the beginning, but very strongly when the mechanical limits are approached. I speculate that there is some kind of sense for costs that may also contribute to the feeling of narrowness. However, these costs may be related not only to the ordinate of Fig. 2a, but also to the abscissa: bodily direction points upward during inhalation and downward during exhalation. Therefore, a factor called cost may be relevant; it is arranged at about 45 degrees relative to the coordinate system of Fig. 2a (see inset, dashed arrow). The assumption that the bodily direction may be correlated with costs is supported by the introspective observation of bodily direction during exhalation and inhalation: The direction appears to be independent of the spatial orientation of the body, be the latter vertical or horizontal. The bodily direction seems rather to be connected to the direction of gravity, against which work has to be done.

One may think of a second factor arranged approximately perpendicular to the cost. This is the intensity of excitation mentioned above. This leads to the idea that, as an alternative to the two introspectively felt dimensions "broad — narrow" and "bodily direction" shown in Fig. 2a, there are two other dimensions rotated by about 45 degrees relative to the first. These new dimensions may be called excitation intensity and costs, respectively, and allow a physiological interpretation. This alternative coordinate system is shown in Fig. 2b. During inhalation, according to Fig. 2a, first breadth increases, then narrowness. The bodily direction goes upwards. During exhalation, the bodily direction is reversed, but, as mentioned by Schmitz, no prominent narrowing can be observed, and only a slight tendency toward broadening. According to the alternative diagram (Fig. 2b), during inhalation, first mainly excitation intensity increases, but later costs increase dramatically. During exhalation, cost and intensity decrease. Introspectively, lowering cost appears to be associated with pleasure and increase of cost (or effort) with displeasure. Thus both terms may be used for this coordinate. These costs are not identical with the physiological costs of a movement, because in joy movement seems easy, but during melancholy the same movement seems to be difficult. As this feeling occurs even before the movement is performed, one should rather speak of imagined (or expected) costs. Both terms, cost and pleasure, are used here as preliminary descriptions only and will be replaced by a third term, namely motor permission, below.

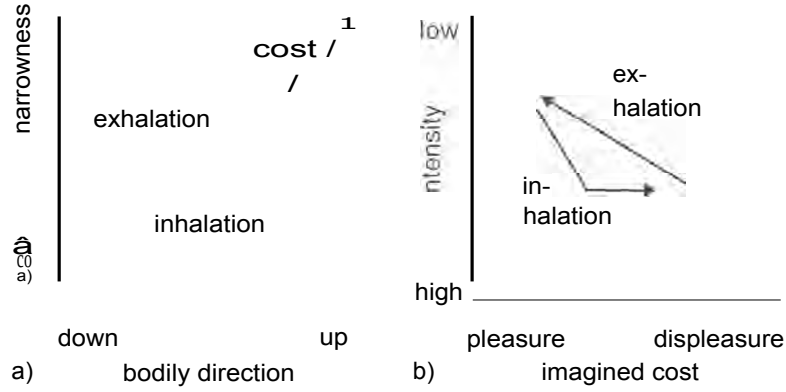

Fig. 2. (a) The coordinate system of H. Schmitz applied to the situation of exhalation and inhalation. (b) shows an alternative system which is rotated by about 45 degrees. The abscissa of (b) is indicated by a dashed arrow in (a).

Fig. 3 shows only the new coordinates with the positions of different (global) emotional feelings and (local) bodily feelings within this system. Is there an interpretation of this coordinate system other than the introspective observations mentioned? For the first dimension (Fig. 2b, ordinate: excitation — intensity), we have already proposed a hypothetical physiological basis in the form of the excitation on a cortical map. How could this second coordinate (Fig. 2b, abscissa) be realized in neuronal terms? We speculate that there exists a "cost map", corresponding to the "intensity map" mentioned above. Why do we require a distributed cost map? Cost could be represented as a single value if we only had to deal with global emotions, like sadness for low pleasure or joy for high pleasure. However, as can be seen in Fig. 3, both dullness and sadness are associated with high costs, or high displeasure. As dullness, like freshness, may be attributed only to parts of the body, costs also require some kind of body representation and are thus assumed to be represented in the form of a map.

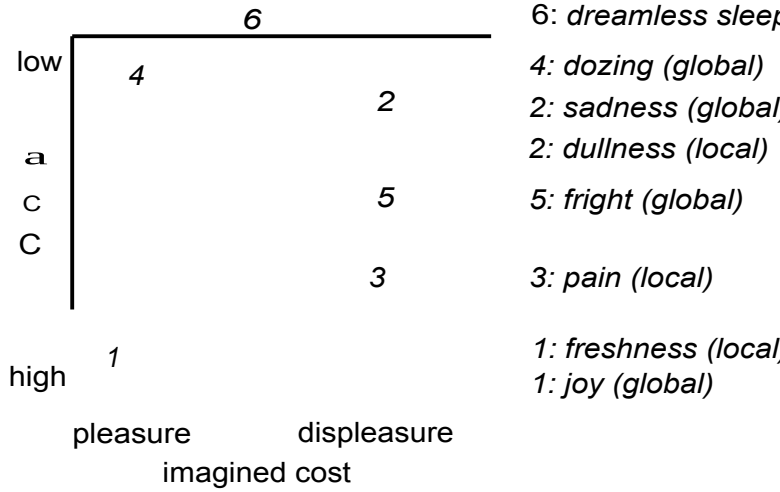

Fig. 3. Coordinate system as shown in Fig. 2b. Arrangement of various emotional feelings (global), and bodily feelings (local) as listed in the right column. Dreamless sleep (6) is related to dozing, but is plotted outside the coordinate systems because it does not correspond to a conscious state.

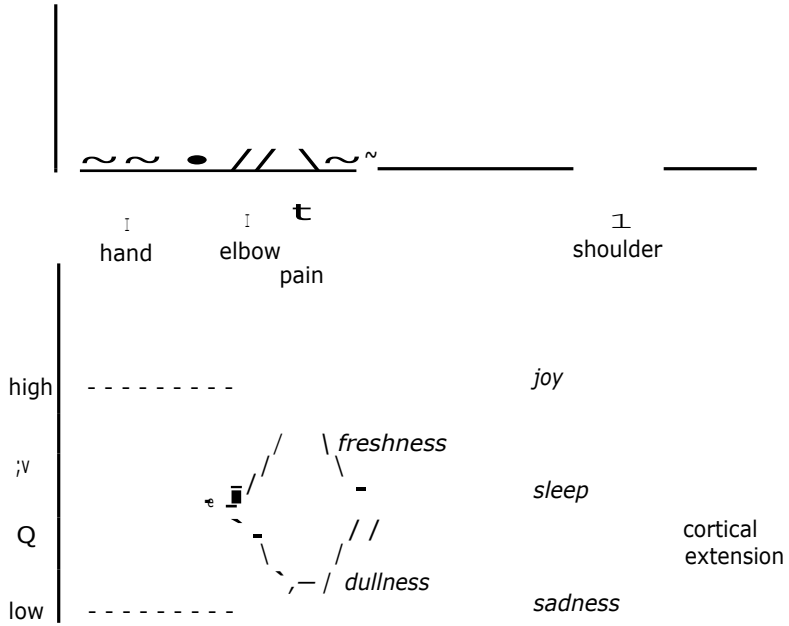

Fig. 4. Two maps. The abscissa in both maps is geometrical extension as used in Fig. 1. Ordinate is sensory excitation intensity in the upper map (intensity) and pleasure (or cost) in the lower map. The latter may also be called motor permission. In the upper map, the dotted excitation may represent concentration of the bodily island on the upper arm. In the lower part, five situations are shown. The arm might feel dull or fresh (local) or the whole atmosphere might be joyful or sad (global). During dreamless sleep, there is neither pleasure nor displeasure. Abbreviations as in Fig. 3.

The two maps postulated here, the intensity map and the cost map, cannot be identical, because high excitation can go together with high costs (in the case of pain) or low costs (in the case of joy). This notion does not, however, exclude the possibility that these two hypothetical "physiological" maps may he morphologically realised in one map.

Compared to Fig. 3, the notation of these two maps illustrates the situation in a different way (Fig. 4). The intensity map is the same as that shown in Fig. 1. The abscissa represents the geometrical extension of this map. The ordinate corresponds to the excitation intensity (see also Figs. 2b, 3; ordinate). Some regions can be excited, but the overall sum of excitation is limited. Furthermore, there is now a second map, plotted parallel to the first, the ordinate of which represents the cost or pleasure (see abscissa of Figs. 2b, 3). The pleasure value could be neutral during sleep; or positive, locally in the case of freshness and globally in the case of joy; or negative, locally for dullness and globally for sadness.

What could be the functional value of the pleasure system? One might speculate that sensory input signalling a positive situation in the environment and producing the feeling of joy, be transformed to allow general influences on the motor system changing the threshold for eliciting a movement. Thus, the pleasure system may make it easier to perform motor actions in positive situations (i.e. those signalled by joy) and preventing their execution in negative situations (i.e. fright, mental or motor fatigue). Such influences may affect the whole body or only selected parts (e.g. dull legs) or they may affect special actions (e.g. flight behavior stimulated by fright). This permissive or inhibitory influence might refer not only to execution of motor actions, but also to the planning of motor actions or to mental activities in general.

Thus, the lower part of Fig. 4 could be interpreted such as describing the subthreshold excitation (corresponding to positive values) or inhibition of premotor neurons influenced by the psychic state of feeling joy or sadness. Therefore, this influence might also be called "motor permission". The higher the "motor permission" value is, the less effort has to be expended to reach the excitation level necessary to execute the movement. This "effort" represents neuronal, "computational" or state costs which have to be distinguished from the "real" costs representing the energy consumption of the muscles. These "task costs" include actions against gravity or the cocontraction of antagonistic muscles often observed in unfamiliar movements. Presumably the sum of both costs is what we experience as pleasure or displeasure, but the second type is not shown in Fig. 4 (lower part).

What might be the functional sense of the intensity map (Fig. 4, upper part)? To speculate how this map might be used to contribute to the

control of behavior (i.e. of motor activity), a more concrete example will be discussed in the following. Let us consider the situation that a person is asked to grasp an object placed on a table in front of him or her. This verbally assigned task must first be transformed into a special motor task like the selection of the appropriate arm and the extension of the joints of the arm. However, before the movement can be started, it has to be decided which other body muscles are necessarily excited or inhibited to allow this reaching movement without violating other conditions, such as keeping the body upright. This requires the application of a body model representing the geometrical (and the dynamical) properties of the body. Such a "mental model" requires as input a goal, e.g. the position of the object to be grasped. Then it can determine which joints have to be moved in which way. Several models have been proposed for this purpose (Kawato 1994, Morass() and Sanguineti 1995, Steinkühler et al. 1995). The latter will be considered in more detail in Section 5 below.

A major problem for the practical application of such a model is how this internal model could be adapted to the geometry of the real body and to its changes (e.g. during growth), in other words, how the internal body model can be learned. Although a number of learning algorithms are known which, in principle, could be applied to train such a neuronal model, the general problem with all these algorithms is that they work well for small systems, but do not easily scale up, due to the combinatorial explosion of the search space. It would thus be an essential advantage if the system could be divided into smaller subsystems, which could then be trained individually. An important question in this context is, of course, how to select these subsystems. Although I do not propose a concrete mechanism of how this selection may be done, I speculate that the local excitation of the intensity map represents the result of this selection procedure. This local excitation may act as a "search light" marking the currently important regions of the map and guiding the internal adapting system to concentrate on these regions.

According to this assumption, excitation in the "intensity layer" might correspond to the selection of attention for learning. This means that those regions which are activated in the intensity layer are switched on for the local updating mechanism of the body model. The whole system would not be learning all the time, but only selected parts.

This fits the following observations. Although one might assume that the sensory information represented in the intensity map is necessary for the control of movement, a series of experimental results shows that a continuous sensory input may not be necessary for the proper performance of an action (Kandel et al. 1991). When, however, the sensory

input is lost for a somewhat extended time, the body model may deteriorate, and performance of voluntary active movements may become worse and, in the extreme case, active movement may no longer be possible (Sacks 1984). Thus, although the sensory input may often not be required to control the actual movement, it may be necessary on the long run because it is needed to keep the body model functioning. Whereas motor permission appears to be more important at the beginning of the behavior under consideration, the information contained in the "sensory" layer may be more relevant after the behavior has been performed.

Up to now we have omitted H. Schmitz' third dimension arranged between epicritical and protopathic. This third dimension might be represented in the shape of the spatial distribution of the sensory excitations over the body map. Epicritic sensation might (hypothetically) correspond to narrow, localized fields of excitation, while protopathic feelings might correlate with widely distributed excitations. This will not be discussed in further details here; nor does H. Schmitz appear to accord these parameters too much importance.

Thus, in summary, I would like to propose that the introspective observations of H. Schmitz may be based on a physiological two-map system, both maps representing the body (the somatosensory map, but also including the other sensory maps and also further regions of more abstract, computational maps). The excitation of the elements of one map parallels some kind of intensity of feeling, those of corresponding positions of the other map parallel the feelings of the imagined (or expected) costs of related movements or actions (Fig. 4). The former is related to the sensory input and is here assumed to support the updating of the body model, the latter is related to the motor system, not in the sense of representing motor commands, but in the sense of judging the cost value of the motor action under consideration. Our subjective experience appears to cover the activities in both maps.

There remain two basic questions. First, what type of neuronal system may underly these "physiological" interpretation of H. Schmitz' system? The second question is even more difficult. Assuming we have found a possible solution for such a neuronal system and are thus able to point to parts of the system whose activities may correlate with the observed psychic states, can we then say anything concerning the basis of this correlation between the "neuronal machinery" and the psychic experiences? In order to discuss the first question (see Section 6), I will expand on the second question first in the two subsequent sections.

4. Epistemological solipsism: the 3rd person's domain is a subset of the 1st person's domain

When discussing the problems related to consciousness, philosophers distinguish between two ontological worlds, which they describe as the external perspective and the internal perspective, or the 3rd person's view and the 1st person's view, respectively. The 3rd person's view concerns our daily intellectual activities dealing with "objective" facts. They are called objective because, in principle, every human has the possibility of mentally treating these objects. Examples are counting the number of apples lying in a dish, or naming the color of a green pencil, or performing a mathematical calculation such as adding two numbers. All this could also be done by a correspondingly programmed robot. These activities are objective or public because they can be looked at by any outside observer. This "external perspective" not only concerns the behavior of the complete system, but, in principle, also includes the possibility of looking inside the brain of a subject and measuring all the interesting neuronal activities, for example, when the subject judges the color of the pencil. In this way, all details of the subject seeing "green" could be determined. These data form the basis of the external perspective, the view of the 3rd person or outside observer. On the other hand, the subject seeing green does not see these neuronal activities, but experiences seeing green. This is called the subjective quality of experience. In contrast to the items of the external perspective, this subjective or 1st person's view is only accessible to the person himself or herself. Therefore, this is also called the private view in contrast to the public view. Nobody other than myself can judge how I see green. The difference between the external perspective and the internal perspective may become even clearer in the example of pain. Again we can consider all neuronal activities that occur when the subject's skin is stimulated, for example by a needle. One might, in principle, even look at one's own action potentials, if oneself is the subject of this experiment. But the experience when considering all these neuronal activities is completely different from the pain one experiences at this moment.

Thus, self observation tells us that there exist systems that can experience an internal perspective. Intuition tells us that there are other systems, like a stone or a simple machine (including some clever nowadays robots) that may not have such an internal perspective.

The assumption that there are two types of systems, some with and others without a 1st person's view leads to the consideration that having the 1st person's view is a system property (sometimes called an

emergent property) that suddenly occurs if some specific conditions are fulfilled. This immediately provokes the question: what are these conditions? Intimately related to this question is the following problem: Assuming we have an idea what these conditions might be, how can we prove that this idea is a sensible hypothesis?

Before I treat these two questions, I will address the above-mentioned distinction between the two worlds determined by the external and the internal view. I would like to argue that this separation is appropriate to describe the historical development and that it is also appropriate as a didactic method to introduce into the problems, but that it is actually misleading when one tries to understand the problems raised here. This separation was probably introduced by Plato and worked out most stringently by Descartes. It can still be found today in most texts refering to these problems. Even in texts tending to oppose Descartes' view, this separation still exists in a more or less hidden form (e.g. Damasio 1994). Unhappy with Descartes' dualistic approach, philosophers have introduced the notion of ontological identity explained with the metaphor of "these are two sides of a coin". This metaphor is definitely helpful in physics in trying to combine two items belonging to the "objective world", for example the wave-corpuscular dichotomy. However, the distinction between the 1st person's view and the 3rd person's view is not cogently described by this metaphor. Looking at the coin requires the capability to look from the outside. This is just what we cannot do from the internal viewpoint. (We do not have the experience, we are the experience: esse est experiri, as stated by Berkeley). I thus propose the following, alternative view which might be called epistemological solipsism. In this view, it is impossible to draw a clear line separating the two worlds in such a way that an item falls into either one or the other domain. Rather, everything of our so-called objective or 3rd person's view is embedded in our subjective experiences. Basically, we have subjective experiences, and those we call objective are subjective experiences that can be more easily separated from each other, therefore being measurable in many cases and more easily transmittable between different persons. Nevertheless, non of these experiences are non-subjective. We do not have access to the "real objective world" (provided such a world exists). Even an "objective" notion like "color is an electromagnetic wave" belongs to the subjective domain, the 1st person's view, because we do not know what an electromagnetic wave really is. A better metaphor would thus be to consider the set of „objective" phenomena as a part, a subset of the subjective world we experience. This epistemological solipsism means that every "objective" phenomenon has some properties that do not belong to all subjective

phenomena. Phenomena lacking these properties are those commonly called subjective. What are these properties? The following Section proposes an answer to this question.

5. A simple neural network model

Let us now come back to the basic questions of how it is possible that some systems have the 1st person's view which Chandler (1995) has termed the really hard problem. What are the conditions allowing humans and, as I believe, "higher" animals, to experience an internal perspective ("have a soul") whereas other systems, like simple animals, dead objects (e.g. a ball rolling in a bowl towards the lowest point, i.e. behaving as if it had a "goal"), or robots, do not experience the 1st person's view. Before I approach this question in Section 6, I would like to introduce three types of systems, which can be distinguished according to their internal architecture (the concepts used in the field of Artificial Neural Networks will be used to describe this architecture). The first type contains a feedforward system, the two other types contain recurrent networks and will be explained below. These descriptions will apply to the 3rd person's view.

A typical feedforward network consists of an input (sensory) layer, one or more hidden layers and an output (motor) layer. Placed in an environment and having the appropriate internal connections, such a system can exhibit quite sophisticated behaviors. The internal connections of this network may be interpreted as comprising an implicit world model (of that part of the world that is of importance for the system), because in some more or less indirect way it represents the properties of the world and the appropriate reactions. This "world model" is static. The reactive or data-driven mode of this system does not allow for explicit predictions in time. To make a prediction possible, additional networks with internal recurrent connections are necessary. Such recurrent networks allow for time-dependent states that are no longer directly dependent on the actual input. In this way, they can comprise world models that can be used by the system to predict sensory input before the corresponding change in the environment actually occurs and to enable the system to prepare the necessary actions accordingly. A simple example is a central pattern generator used to control time-dependent behavior. This not only concerns the control of the motor output, but also the time-dependent control of sensory input. Usually these "world models" are confined to a given behavioral context elicited by a definite complex of sensory (or internal, e.g. determined by the

hormonal state) stimuli, and, in this respect, are also, however, more indirectly, data-driven. These two types of systems will be described as containing non-manipulable models, in contrast to a third type of world model to be explained below. Both types of systems described up to now are intuitively assumed to have no 1st person's view.

As a further step, one could now imagine expanding the system such that a third type of internal world model exists that can be to a large degree uncoupled from the actual sensory input and the motor output. For example, this is the case when a movement is planned in order to decide whether this movement, when executed, would lead to the desired result. This movement may, for instance, concern the grasping of an apple lying on the table possibly within reach, or, to take a more abstract example, the movement of a figure in a chess game. Such an internal world model may be called manipulable, because after the goal is given as input to this model, it can be further uncoupled from actual sensory input and "can play around in its virtual internal world". The results this model proposes after finishing its search can be judged by given evaluation criteria. Depending on these values, a decision might be made whether or not to actually perform the proposed activity.

As long as this world model represents only information concerning the outer world and not properties of the system itself, one still might not be prepared to attribute a 1st person's view to it. (Traditional, non-situated AI systems may be considered examples of such cases). This situation might change if the system under view, first, has not only sensory but also motor contact with the physical world and, second, some of the physical properties of the system itself are also embedded into the internal world model. To explain this, 1 would like to restrict the following discussion to the experiences of the subject's own body including planning and performing movements, for example reaching movements with the arm and the hand. I do this, first, because a concrete example is clearer than general and abstract levels of argument and, second, the body model that animals including humans use when planning or performing a movement may be the very basis of the phylogenetic evolution and possibly also the ontogenetic development of a manipulable internal model. However, it should be mentioned that consideration of the task of reaching for an object, e.g. a glass, also requires expanding the model to include such external objects. The most primitive "cognitive" system may have only knowledge concerning its own body. To solve such a reaching task, however, this model has to be enlarged by the capability of embedding objects which occur in the workspace. Furthermore, more distant objects could be integrated to permit, for example, orientation. Even at this stage of development, this model

contains internal representations or "concepts" of these objects, although application of language is not yet considered. The invention of spoken words by humans increased the manipulability of these virtual objects, allows for finer distinction between items, and improves the possibility of communicating between subjects. At least on this level, items stored in the manipulable model can be called "objective". In this way, the internal world model starting with representations of only its own body is expanded to contain near and distant extracorporeal objects as well as abstract entities. But all these are, I assume, basically related to the representation of the own body.

As before, the following description of the system will start using the 3rd person's view. In the system discussed now, we have two representations of the world. The first one corresponds to the two non-manipulable models described above, which are given by the current sensory data and by stored data earlier obtained from proprioceptors, but also from exteroceptors, for example the visual system observing the subject's own body (or registering reactions of the environment to an action of the subject). In this way, a static and/or dynamic, but non-manipulable model of the world, including information concerning the subject's own body can be constructed. The second set of representations is given by the output of the manipulable model.

To explain the hypothesis more concretely, in the following an extremely basic model will be described, containing a non-manipulable and a manipulable model. As mentioned, Steinkühler et al. (1995) proposed a model that can deal with the problem of inverse kinematics, i.e. that can determine a geometrical arrangement of a multilimbed arm to grasp a given object. This model provides a holistic internal model of the kinematics of the body. Although the model can be extended to describe the geometry of a complete body (Cruse et al. 1996), I will concentrate here on the example of moving a three-joint planar arm. The control of such an arm **already** involves the general problem of controlling a redundant system, because the end point of the planar arm is determined by two coordinates, whereas more than two independent values, namely the three joint angles, have to be controlled. This "redundant" system thus has one extra degree of freedom. The model consists of a recurrent network that relaxes to adopt a stable state corresponding to a geometrically correct solution, even when the input does not fully constrain the solution. The underlying idea is that several geometrical relationships are used to calculate the same value (e.g. the angle of the elbow joint) several times in parallel. The final output is then obtained by calculation of the mean value of these multiple computations (MMC). This is then fed back for the next iteration. Any

special constraint of the variables (e.g. joint limits) can be introduced. In this way, the MMC-type network can be used as a kinematic inverse model which allows for very different solutions of the redundancy problem and which can easily cope with limitations of the joint space or the workspace.

The MMC model will be shown in a simplified form. This presentation permits an explanation of the MMC system as a combination of a recurrent system and a simple multilayered feedforward system. The latter is called the forward model in Fig. 5. The sensory variables are called $a_{5,}$ $_{ßs,}$ and y_s, describing the joint angles of, say, shoulder, elbow, and wrist. These inputs are used to reset the recurrent connections of the internal model before this can be used. The position of the end point of the arm is defined **here** by the Cartesian coordinates x,y given by the visual system. The output is given by a "motor layer" determining the joint positions a_m, 13_m, and y_m. This internal model can be used to plan the movement (or to be considered as the neural basis for the production of a mental image of an imagined movement which has been called *Bewegungsphantasie* by Gehlen, 1971) or to control an actual movement. In the first case, the connections to the motor layer are interrupted by an influence called PLAN in Fig. 5. This influence may also inhibit the resetting. There are reasons to believe that not all movements are controlled by such a body model. On the contrary, for urgent actions there are presumably also many parallel systems for direct specifications of actions. These systems, however, are normally not cognitively penetrable and are therefore called non-manipulable)`. They are assumed here to correspond to the non-manipulable models and are summarized by the box DIRECT in Fig. 5.

It might well be that these directly specified action subsystems are also constructed based on this MMC principle without, however, a switch corresponding to the PLAN unit (instead, the decision between these action subsystems may be controlled by an additional motivation layer as shown by Maes, 1991). Using the MMC principle to store, for example, the item "house" would mean that the MMC net, in a holistic manner, contains the information about the visual input corresponding to the view of a house, the acoustic input of hearing the word "house", the output of uttering "house" or for example of drawing a house.

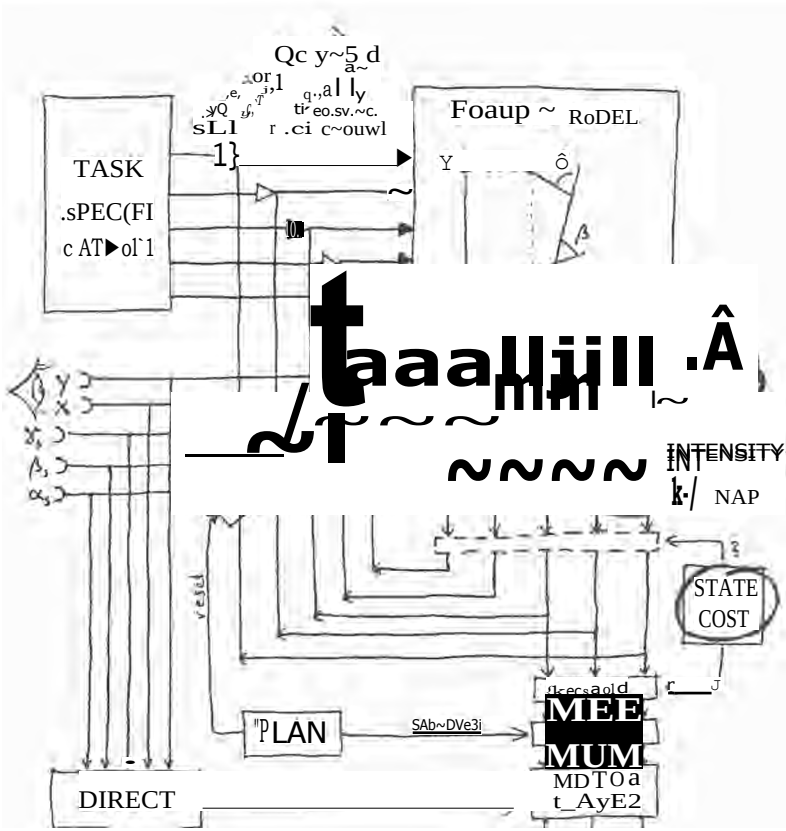

Fig. 5 Preliminary sketch showing an expanded version of the MMC model for a three-joint, planar arm. Parts marked by ellipses indicate the information that may contribute to feelings. PLAN: inhibits the motor output and the resetting of the recurrent connections during "planning". For further explanations see text.

As discussed by Steinkühler et al. (1995), the MMC model could be extended to include additional conditions reflecting cost values, e.g. the static costs depending on a joint position deviating from a comfortable medium position. Similarly, it could also be expanded to include costs arising from cocontraction of antagonistic muscles controlling the same joint. In the latter case, the sum of these excitations corresponds to the costs of holding the desired position. These could be used as an extension of the model (Fig. 5, task costs) to find a solution that minimizes the costs to reach a given goal (e.g. a desired position of the hand). These task-dependent costs most probably have to include dynamic costs that depend on the initial position and the desired speed, which, however will not be further considered here.

As mentioned above, beside the task costs, there is a second type of costs which do not depend on the task itself, but on the psychic state of the person. How could these state costs influence the system? One possibility is that during joy or depression the subthreshold inhibition of the (pre)motor neurons might be low or high, respectively. These inhibitory signals can be regarded as (psychic state-dependent) costs. The state costs influence might also affect the recurrent connections, such that the system relaxes faster or slower. The latter possibility is indicated in Fig. 5 by a questionmark. What seems to be introspectively experienced when a movement is being planned appears to correspond to some kind of sum of the task costs and the state costs. Both are indicated by ellipses in Fig. 5. Why this leads to an experience is an open question (for some discussion see below, Section 6)

Another, simpler open question is how the body model can adapt to changes of body geometry, e.g. during growth, or to sensory drift. This could be done in the following way. When the difference (shown by circles in Fig. 5) between the output of the body model, e.g. joint angles, and the actual sensory input is zero, the body model is appropriate. In this case no changes of the internal connections, the synaptic weights, are necessary. The greater the differences, the greater the necessary changes of the weights. Using these differences as error signals (Fig. 5, dashed lines), the body model could easily be updated using the classical backpropagation algorithm in accordance to a proposal by Kawato and Gomi (1992).

According to the above hypothesis (see Section 3), it appears to be sensible that only a limited part of the complete internal model is selected for learning. As mentioned in the example of Fig. 5, the sensory map is restricted to the three angles of shoulder (a_s), elbow (P_s) and wrist (y_s), and the coordinates of the grasp space x,y to a 2D plane. Therefore, in this particular example, because of the already small

number of parameters a further selection may not seem very sensible. On a higher level, this selection probably not only concerns the body parts as such, but also the general behavioral context (e.g. whether the muscles are used during cycling or skiing). This selection is assumed to occur at the level of the recurrent connections of the model. It should be mentioned that, because of the holistic nature of the model, these connections can be labeled to carry both motor or sensory values. There is no distinction possible at this level. As will be described in more detail in the following section, I assume that this mechanism of selection results in the body sections actually being felt (Fig. 5, intensity map). The application of H. Schmitz's observation to this model leads to the hypothesis that the activity within the recurrent layer correlates with the intensity layer (Fig. 2b) and thus with a combination of H. Schmitz's parameters narrowness/breadth and bodily direction (Fig. 2a). Can we say anything concerning the correlation between the "neuronal machinery" and psychic experiences?

6. Conditions for the appearance of the 1st person's view

Apart from enabling us to learn the properties of the body model, this comparison between the output of the body model and the actual sensory input might also have another effect. When the manipulable model proposes a new arm position, the latter may be different from the actual one represented by sensory input. Functionally, this difference may be used to control a corresponding movement and this may be a sufficient explanation when applying the 3rd person's view. At this point, however, let us also consider the 1st person's view. When planning a movement we experience a body image, we "see" — or can imagine — the position of our virtual arm moving around. Why does this experience occur? Seen by an outside observer, the situation can be described such that two data sets exist that both represent the subject's own body plus related associations. One is represented by the current activities of the actual sensory input and the stored memories, the other by that of the manipulable model. The data of both representations coincide when the person is in a mentally, psychically relaxed state; they may differ when the person imagines a movement or when he or she actually performs a movement and, because of its inertia (for example), the arm drags behind the imagined position. I assume that it is the confrontation of these two data

sets which yields the condition for the system to experience a 1st person's view (see Cruse 1979; for an excellent and very elaborated description of this view, see Metzinger 1993). Because there is no outside observer, the hypothesis proposed here is that it is this *difference* between the two representations that is experienced, and which makes up the 1st person's view: therefore Berkeley's statement should be changed to esse est experiri differentiae. To illustrate this, in Fig. 5 the ellipse called "intensity map" includes those parts of the model where these differences are determined

I do not see a way to make it intuitively clear that, under the conditions described, a 1st person's view necessarily develops. However, this is similar in other cases where unexpected system properties emerge as, for example, the property of resonance frequency in an electronic circuit. After having observed and described the resonance phenomena, including constructing a pendulum-like mental model, we have the impression of having understood the property of resonance. Another, possibly even better example concerns the occurrence of the phenomenon of life. After a number of investigations, we have developed a mental model stipulating that a special collection and arrangement of molecules shows the property of living.

The hypothesis is that it is the occurrence of this difference that makes the system feel. This hypothesis may be of interest if, using the external perspective description, properties of the system are found that correspond to observations attributed to the 1st person's view. Two such properties will be explained. As mentioned, according to our hypothesis only the differences can be recognized, but not these two representations as such. The latter can only be described using the 3rd person's view. The 1st person's view sees only the difference. This agrees with the introspective observation that the conscious I does not have the impression of viewing a movie (like a homunculus would), but to be in the world. This is sometimes called transparency (Metzinger 1993). Because the basis of the recognized difference is the subject's own sensory and memory data, the I is always in the center of the experience (feeling) or as David Bell stated, the I is immune to error through misidentification; philosophers call this perspectivity (Metzinger 1993). This provides the impression of the (subjectively) constant I, although, seen from the 3rd person's view, we actually do change a lot during our lifetime.

The exact border dividing observations belonging to the I and belonging to the environment is not strictly fixed. A hammer in the hand can be experienced in the body image as an extension of the body (H. Schmitz calls this *Einleibung).* Also, a car driver may feel his body extended to the outer margin of the car's body. One may even feel kind

of sick when the car does not work well. Depending on cultural influences, these borderlines may even include other people; an extreme example, for members of the western culture, may be the experience of labor-pains by males in the case of couvade.

Another aspect concerns the ontogenetic development of the property having a 1st person's view. Our intuition tells us that a human foetus in a very early state, say in the 64-cell stage, does not have a 1st person's view. But at what stage of development does this appear? According to Piaget (1963), in infants the separation between the I and the rest of the world occurs during the second year. Before this time, the world model may not be separable from the sensory data, and therefore not manipulable in the above sense. This leads to the hypothesis that infants develop the 1st person's view around this age at the latest.

Similar experiences have been reported when an accident impairs brain structures destroying the consciousness of the existence of e.g. a leg as being a part of the body. At the same time, the leg cannot voluntarily be moved, although the muscles and peripheral nerves are intact (Sacks 1984). Regaining motor control parallels the experiencing of that leg. This suggests that only such items have the ability to become experienced that are represented in the manipulable world model.

Taken together, these observations support the following hypothesis: a system has a 1st person's view, when (a) it contains a manipulable internal world model that includes properties of the system's own body which (b) can be used to compare the ("virtual") data of this model with those provided by the "real" data from the sensory input and the memory. As a weaker form of the hypothesis, one can assume that these conditions are necessary for the 1st person's view to appear as an emergent property. The stronger version assumes that these conditions are also sufficient.

7. Discussion

One may not be intuitively prepared to attribute a 1st person's view to the simple system shown in Fig. 5a. However, most probably each concrete proposal — i.e. realisable in hardware — for an expansion of the system would lead to the impression of a system not having a 1st person's view. Because, however, such systems do exist, this observation implies that we would not recognize the transition from a system lacking to a system having a 1st person's view, even if it had happened. Therefore, intuition is a problematic advisor and, using Occam's razor, we should try to start with the simplest explanation. Another argument to

help intuition might be the following. Such a simple system may indeed have crossed the principal threshold, but the liveliness of its 1st person's view might actually be very weak. The liveliness might become more intense and the feeling of the system may become more and more real, the more different sensory modalities contribute to the description of a given situation. This is supported by the observation of patients with sensory deficiencies who report that their feeling seems to be less real. In this way, a gradual change is conceivable, ranging from a very weak 1st person's view to full awareness, a change which we also experience during our daily life.

Is there a possibility to test this hypothesis? The answer is that there is probably no direct possibility to test it. The only immediate way to prove the feasability of a hypothesis concerning the 3rd person's view is, as mentioned above, to collect observations on the 1st person's view and see whether these observations could be translated to the 3rd person's view in the sense that they appear to corroborate the hypothesis. At first sight, a direct proof might be completely rejected. This is the idea behind Nagel's question: What is it like to be a bat? To answer this question, one would first need the sensory experience and the memory of a bat and, second, the manipulable world model of the bat in order to be able to compare them. In short, this is possible only for a bat. However, as a gedankenexperiment, there are ways of improving the intuitive acceptability of this hypothesis. One way is to built an exact copy of a human so that one can assume that this „machine" has a 1st person's view. Then, neglecting moral problems with respect to this machine, one could, step by step, change this technical system to make it more and more similar to a bat. In this way, at least this humanoid system could "slip" into the way of being a bat. If it retained its ability to speak, it might even be able to report its experiences during this journey.

Finally, two minor points should be mentioned. First, the two-map arrangement shown in Fig. 4 could be interpreted to mean that the morphological basis of the motor permission map corresponds to the premotor area and the intensity map to the secondary somatosensory and other higher-order sensory areas. However, too little is known about the functional properties of the brain structures to make any sensible statement. Of course, various subcortical structures might contribute or may even be essential elements. As mentioned, it does not necessarily follow that these two maps are morphologically separated. Simple architectures are conceivable that would merge both physiological functions into one morphological layer.

Second, it was already mentioned that H. Schmitz posits an antagonistic relation between the two tendencies narrowing and broadening.

He distinguishes between two forms of antagonistic behavior, what he calls simultaneous coupling and rhythmic coupling. During simultaneous coupling, both tendencies influence each other in a way leading to a static or only slowly-changing situation in which one or the other tendency dominates. Examples are pain or inhalation. The other form corresponds to the cooperation of dynamic, for example pendulum-like systems, in which the mutual influence of these tendencies leads to a rhythmic change such that narrowing and broadening dominate alternately. According to H. Schmitz,. this is the case for lust, tickling, or anxiety. Although a very important aspect, this dynamic type of cooperation has not been considered sufficiently here.

References

Aubè, M., Senteni, A. (1996): "Commitments management and regulation within animals/animats encounters". In: Maes, P., Mataric, M.J., Meyer, J.-A., Pollack, J., Wilson, S.W. (eds.): *From animals to animats. 4. Proceed. of the Fourth Intern. Conf on Simulation of Adaptive Behavior*. 264-271, MIT Press, Cambridge MA.

Bell, D. (1997): Solipsism and Subjectivity. Paper presented at the Wissenschaftskolleg 1996. (forthcoming)

Cabanac, M. (1992): "Pleasure: the Common Currency". *J. Theor. Biol.* 155, 173-200.

Chandler, D.J. (1995): "The Puzzle of Conscious Experience". *Sci.Am.* Dec. 62-68.

Cruse, H. (1979): "Modellvorstellungen zu Bewußtseinsvorgängen". *Naturw. Rundschau* 32, 45-54.

Cruse, H., Bartling, C., Dean, J., Kindermann, T., Schmitz, J., Schumm, M., Wagner, H. (1996): "Simple solutions to complex problems by exploitation of the physical properties". In: Maes, P., Mataric, M.J., Meyer, J.-A., Pollack, J., Wilson, S.W. (eds.): *From animals to animats. 4. Proceed. of the Fourth Intern. Conf. on Simulation of Adaptive Behavior*. 84-93, MIT Press, Cambridge MA.

Damasio (1994): *Descartes' `Error'. Emotion, Reason and the Human Brain. G.P.* Putnam's Son, New York.

Gehlen, A. (1971): *Der Mensch.* Athenäum, Frankfurt.

Kandel, E.R., Schwartz, J.H., Jessell, T.M. (1991): *Principles of Neural Science.* Appleton & Lange.

Kandel, E.R., Schwartz, J.H., Jessell, T.M. (1995): *Essentials of Neural Science and Behavior.* Appleton & Lange.

Kawato, M. (1994): "A Bi-directional Theory Approach to Prerational Intelligence". In: *ZiF Preprint Series. Prerational Intelligence in Robotics: From Sensorimotor Intelligence to Collective Behavior.* Bielefeld.

Kawato, M., Gomi, H. (1992): "The Cerebellum and VOR/OKR Learning Model". *Trends in Neurosciences* 15, 455-453.

Maes, P. (1991): "A bottom-up mechanism for behavior selection in an artificial creature". In: *From animals to animats.* J.A. Meyer, S.W. Wilson (eds.). pp. 238-246. MIT Press, Cambridge, MA.

Metzinger, T. (1993): *Subjekt und Selbstmodell. E* Schöningh, Paderborn, München, Wien, Zürich.

Morasso, P., Sanguineti, V. (1995): *Self-organizing Body-schema for Motor Planning.* J. Motor Behavior 27, 52.

Piaget, J. (1963): *The Origin of Intelligence in Children.* Norton, New York.
Sacks, O. (1984): *A Leg to Stand on.* Gerald Duckworth, London.
Schmitz, H. (1965): *System der Philiosophie. Der Leib im Spiegel der Kunst. Vol II,* part 1, pp. 73-172 Bouvier, Bonn.
Schmitz, H. (1966): *System der Philiosophie. Der Leib im Spiegel der Kunst. Vol II,* part 2, pp. 7-36. Bouvier, Bonn.
Schmitz, H. (1969): *System der Philiosophie. Der Gefühlsraum. Vol III,* part 2, pp. 91-401. Bouvier, Bonn.
Schmitz, H. (1989): *Leib und Gefühl.* Junfermann Paderborn.
Steinkühler, U., Beyn, W-J., Cruse, H. (1995): "A Simplified MMC Model for the Control of an Arm with Redundant Degrees of Freedom". *Neural Processing Letters* 2, 11-15.