Available online at www.sciencedirect.com
SCIENCE DIRECT®
Consciousness and Cognition 12 (2003) 515–528
Consciousness and Cognition
www.elsevier.com/locate/concog

Emerging selves: Representational foundations of subjectivity☆

Wolfgang Prinz

Max Planck Institute for Psychological Research, Amalienstrasse 33, D-80799 Munich, Germany

Received 27 February 2003

Abstract

A hypothetical evolutionary scenario is offered meant to account for the emergence of mental selves. According to the scenario, mental selves are constructed to solve a source-attribution problem. They emerge when internally generated mental contents (e.g., thoughts and goals) are treated like messages arising from external personal sources. As a result, mental contents becomes attributed to the self as an internal personal source. According to this view, subjectivity is construed outward-in, that is, one's own mental self is derived from, and is secondary to, the mental selves perceived in others. The social construction of subjectivity and selfhood relies on, and is maintained in, various discourses on subjectivity.
© 2003 Elsevier Inc. Open access under CC BY-NC-ND license.

1. Introduction

This paper offers an account of how mental selves may have emerged. It comes as a piece of speculative psycho-history, that is, as a hypothetical evolutionary scenario. The core message is that mental selves are sociocultural constructs rather than naturally given organs of human minds. My argument will take three steps. First, I shall point out how the emergence of conscious awareness is linked to the emergence of selves. Then, I shall offer an evolutionary scenario that explains how, at some point far back in Stone Age, mental selves came into being as a solution to a particular representational problem. Finally, I shall discuss mechanisms of social interaction and discourse in the service of maintaining and preserving the institution of mental selves.

☆ I am grateful to Jonathan Harrow for translating the text from German into English; Heide John for completing the manuscript.

E-mail address: prinz@psy.mpg.de (W. Prinz).

1053-8100 © 2003 Elsevier Inc. Open access under CC BY-NC-ND license.
doi:10.1016/S1053-8100(03)00053-9

2. Conscious experience and subjectivity

When we talk about stars and stones, when we are dealing with algae and fungi, or even grasses and trees, we generally believe that everything that can be said about them at all can be said through the means of physics, chemistry, and biology. This may also still hold for jellyfish, worms, and sponges, but then things soon start to become critical—if not with insects, then perhaps with vertebrates, and certainly with mammals, particularly the primates, and inevitably, of course, with *Homo sapiens*—with you and me. When considering these animals, we believe we cannot say everything that can be said about them with these sciences. We also attribute them with an *inner life*, perhaps to a lesser degree, but, in principle, similar and related to our own conscious experience. Some believe that when we impute subjectivity, we attribute to others what we know from ourselves. Some others believe that we attribute subjectivity as a theoretical construct that applies to others as well as to ourselves.

Subjectivity or consciousness—how should one approach this new quality that seems to have emerged somewhere between fish and *H. sapiens*? How does subjectivity enter the world? What is its function?

2.1. Descriptive accounts

What does conscious experience really mean? Let me start by stating what I do *not* mean when I use this term. First, I do not mean those concepts of consciousness that take the rank of a special substance, ontological form, or ontological level, as found, for example, in the German school of idealism; instead, I use the concept and its verbal derivatives exclusively in the *psychological sense*, that is, in the sense of concrete forms of mental experience. Second, I do not mean concepts of consciousness that describe certain states of persons, for example, that someone is conscious or has lost consciousness and is then unconscious. This notion of consciousness describes the *disposition* to produce conscious mental contents, and, insofar as theories exist on this, they specify only the conditions under which this disposition comes and goes without telling us anything about what is responsible for the conscious nature of their content.

This brings me to what I do mean and what I wish to explicate here: *the conscious nature of mental contents*. Our being conscious is always expressed by our being aware of *specific contents*. We cannot be conscious in a pure or empty form. It is always *specific contents* of which we are aware, and it is only through our awareness of this contents that we recognize that *we are conscious*. Thus, like persons who can be in conscious or unconscious states, mental contents can be conscious or nonconscious as well. Depending on our theoretical tastes, we then use such labels as *unconscious* or *preconscious* mental contents, or those that are *inaccessible to conscious experience*.

How, then, is *conscious experience* possible, and what exactly is its conscious nature? When answering this question, I shall follow a characterization given by Brentano (1924). Brentano uses an extremely simple example to discuss the nature of what he calls mental acts. He raises the question what actually happens when we hear a tone. What is responsible for the conscious nature of this event? According to Brentano, there are two contents interwoven in this mental act: the *tone* that we hear and the fact that we *hear* it. However, these two mental contents are not represented in the same way. The tone is the primary object of hearing; we can observe it directly in the mental act. Hearing itself is the secondary object of the mental act. Brentano says that it cannot be observed directly, but

attains consciousness in another, indirect form: "We can observe the tones that we hear, but we cannot observe the hearing of the tones. This is because it is only in the hearing of the *tones that the hearing* itself is also assessed" (Brentano, 1924, p. 181; italics added).

This is as far as Brentano goes. However, if we want an exhaustive characterization of the structure of mental acts, we have to go one step further: If it is true that hearing the tone contains not only the tone itself but also, implicitly, its hearing, then the subject hearing it must also be contained in the act in another encapsulation. This is because, just as a tone is hardly conceivable without a hearing directed toward it, a hearing is hardly conceivable without a mental self or subject who hears. Hence, conscious mental acts are characterized by what is sometimes termed "me-ness" (e.g., Kihlstrom, 1997), that is, the implicit presence of the mental self.[1]

2.2. Explanatory accounts

What sort of explanations do we need in order to finally say that we *understand* how the conscious nature of mental contents arises? If it is true that the relationship of mental contents to an implicitly present self forms the decisive foundation for the formation of their conscious nature, then the problem of explanation shifts from consciousness to the self, that is, to the issue of the constitution of the mental self and its implicit representation.

Therefore, we need theories that explain the role of the implicitly present mental self. An understanding of this role would simultaneously mean an understanding of how the conscious nature of mental contents arises. Because the quality of conscious awareness does not just *arise when* the condition of the implicit presence of the self is fulfilled but also *consists precisely* in this condition being met, it becomes possible to understand not only—in correlational terms—*under which conditions* conscious mental contents emerge but also—in foundational terms—*why* they assume *precisely this quality* and not any other.[2]

[1] In line with this, the conscious representation of a situation ends precisely when the self departs from it. If, e.g., while out walking, we become involved in a conversation that claims our entire attention, our conscious perception is focused on the content of the conversation and the conversational situation itself. These are, to return to Brentano, the primary objects on which the implicitly present self concentrates. It is only these that we perceive consciously. Other features of the situation—namely, the scenery we are passing through—are not acknowledged. Naturally, there can be no doubt that this information is *processed*, because, otherwise, it would be impossible to explain why we are completely able to adjust our steps in line with the environment, even when we are deep in conversation. However, this processing does not generate any *conscious* representation, representation related to the implicitly present self.

[2] This distinction addresses one of the major problems faced by research into what some have termed the NCC—the *neural correlates of consciousness* (e.g., Metzinger, 2000): Even if we had an exact and comprehensive account of the neural correlates of consciousness, we would still be far from understanding why these neural processes bring forth precisely this particular quality, and not some other. And, vice versa, we would be just as uncertain why certain neural processes bring forth the quality of conscious awareness whereas others, in contrast, do not. Theories on the relation between brain functions and consciousness offer, at best, *correlational* relations that help us to *know* how the brain subserves conscious experience, but not *foundational* relations that help us to *understand* those relationships. In this sense the chase for the NCC appears to be pointless. The disparity between the phenomena we wish to relate to each other here seems to be too large to permit us to trace them back to each other directly. Therefore, in order to gain a better understanding of the foundational relations we are seeking, we need to insert a further intermediate level of description *between* brain processes and conscious experience, at which the foundations of conscious experience are specified at a *representational level*. This is exactly what this evolutionary scenario aims at.

What do we mean by explaining the role of the mental self? What do we mean in any case by *explaining a role*? We need to explain *what* it performs on an ultimate level and *how* it achieves this performance on a proximate level. Only when we have both—performances and mechanisms—can we really understand the conscious nature of mental contents.

What, then, may mental selves be good for and why have they emerged during evolution (or, perhaps, human evolution or even early human history)? Answers to these questions used to take the form of stories explaining how the mental self came about and what advantages were associated with it. In other words, these are theories that construct hypothetical scenarios offering plausible explanations for why certain (groups of) living things that initially do not possess a mental self gain fitness advantages when they develop such an entity—with the consequence that they move from what we can call a self-less to a self-based or "self-morphic" state.

Modules for such scenarios have been presented occasionally in recent years by, for example, Dennett (1990, 1992), Donald (2001), Edelman (1989), Jaynes (1976), Metzinger (1993, 2003), or Mithen (1996). Despite all the differences in their approaches, they converge around a few interesting points. *First,* they believe that the transition between the self-less and self-morphic state occurred at some stage during the course of human history—and not before. *Second*, they emphasize the cognitive and dynamic advantages accompanying the formation of a mental self. And, *third*, they also discuss the social and political conditions that promote or hinder the constitution of this self-morphic state. In the scenario below, I want to show how these modules can be keyed together to form a coherent construction.

3. The emergence of selves: A stone-age tale

3.1. Two basic principles

As a starting point for the evolutionary scenario, let us consider a fictitious, highly organized living thing—on approximately the organizational level of a higher mammal. Although this animal is equipped with highly developed cognitive abilities, it has one typical limitation: It is a *preverbal* being in which symbolic communication and representation do not occur.

The cognitive ability of this fictitious animal, but also the limits to its performance, can be sketched roughly as follows: On the plus side, we can attribute it—put briefly—with the ability to evaluate the behaviorally relevant implications of each current stimulus situation and to transform these evaluations into appropriate behavior. This evaluation functions on the basis of complex algorithms that may belong, in part, to its genetically determined behavioral repertoire but may also have arisen through learning processes. Further algorithms ensure that the outcomes of these evaluations are compared with the animal's current needs, and that this comparison is transformed into action decisions.

Regardless of the potential complexity of the computations underlying behavior control, they remain subject to the principal constraint of being *coupled with each current situation.* They commence with the current stimulus information and evaluate action options that refer to the current situation. In contrast, processes requiring explicit representation of circumstances that are not currently present, such as representing past or planning future events, are not involved at all. Our model animal is chained to the present.

This hypothetical basis forms the background for the following postulate that will be elaborated below: From these initially self-less beginnings, self-morphic modes of representation may emerge, and can only emerge when two developmental stages follow each other in sequence. At the first stage, it is necessary to develop the ability to re-present circumstances that are not currently present and keep such re-presentation separate from perception.[3] In the following, I shall label this precondition *dual representation.* At the second stage, this has to be joined by representation being interpreted as resulting from personal communication. I shall label this precondition *attribution to persons*.

Dual representation concerns the *natural history* of behavior organization; attribution to persons, in contrast, concerns the *cultural history* of our species. Although these two lines of development can be distinguished systematically, they are linked together so closely in historical terms that they can be presented only in relation to each other. I shall now describe this relationship for two distinct mental domains: first, for cognition and, second, for action.

3.2. Application cognition: Thoughts and their sources

Dual representation. We shall now extend the abilities of our fictitious animal through one decisive step and assume that the social association in which it lives develops simple forms of symbolic communication. What does this require? Let us examine, for example, the case of a message referring to a circumstance beyond the current perceptual horizon of its recipient. To be able to understand such a message, our animal must possess the ability of re-presentation, that is, to form mental re-presentations of circumstances that cannot be perceived at the present time. When doing this, it has to be able to discriminate between the contents of perception and re-presentation, because it has to ensure that this re-presentation does not impede actions in the current perceptual situation.

The formation of re-presentation therefore has two sides: On the one side, it permits a representational decoupling from the current situation. On the other side, however, re-presentation cannot be allowed to replace perceptions; these have to be retained completely in order to guide action. The simultaneous processing of re-presented contents *alongside* perceived contents calls for a far-reaching extension of the cognitive processing architecture. What is needed now is an architecture that discriminates between foreground and background processing and makes it possible to process re-presented information in the foreground while simultaneously continuing to process current perceptual information in the background—at least to an extent that maintains the elementary functions such as movement control or orientation reactions in response to unexpected stimuli.

[3] The term *re-presentation* stands for the German term *Vergegenwärtigung* that covers all kinds of representations referring to circumstances that are not present at the current time. In the following, I shall distinguish between re-presentations (in this restricted sense) and representations (in the general, broader sense). The term *dual representation* refers to the requirement to keep *re-presentation* (of what is absent) and *perception* (of what is present) separate from each other.—Importantly, the concept of *re-presentation* must not be confused with the notion of *re-representation* recently suggested by Whiten, which refers to a specific way in which an individual interprets another individual's actions (Whiten, 2000).

The mode of representation coupled with this new organization of information processing is what I label dual representation. I understand this as the ability to run perceived contents and re-presented contents *in juxtaposition* but *separated functionally*. I do not wish to speculate about how this architecture is implemented in the brain. The only important thing here is that it extends the cognitive organization potential of the animal equipped with it in many ways. The most important extension is probably the *birth of the self* as a consequence of attribution to persons.[4]

Attribution of thoughts to persons. Up to now, we have considered only re-presentation that is triggered by the reception of verbal messages and, as such, is induced from the outside. Once an architecture of dual representation has been formed, it also offers space for the induction of re-presentations from within such as thoughts, memories, or fantasies. For the sake of brevity, I shall use the expression *thoughts* to stand for all forms of internally induced mental re-presentation.

Internally generated thoughts differ from externally induced messages in one important aspect, though: When re-presentation is triggered by external verbal messages, it is always accompanied by the *perception of an act of communication* that itself occurs in the current perceptual situation. In other words, there is always a person in the receiver's environment, and this person is the perceivable source of the message. *Thoughts*, in contrast, are internally generated acts of re-presentation that are not accompanied by the perception of a current act of communication—so that they cannot be attributed to any external human source in the current situation.

Thus, where do thoughts come from? Who or what generates them, and how are they linked to the current perceptual situation? This brings us to a problem that psychology describes as the *problem of source attribution* (Heider, 1958).

One obvious suggestion is to transfer the schema for interpreting externally induced messages to internally induced thoughts as well. Accordingly, thoughts are also traced back to human sources and, likewise, to sources that are present in the current situation. Such sources can be construed in completely different ways. One solution is to trace the occurrence of thoughts back to *voices*—the voices of gods, priests, kings, or ancestors, in other words, personal authorities that are believed to have an invisible presence in the current situation. Another solution is to locate the source of thoughts in an autonomous personal authority bound to the body of the actor: the self.

These two solutions to the attribution problem differ in many ways: historically, politically, and psychologically. In historical terms, the former must be markedly older than the latter. The

[4] The assumption that dual representation is a prerequisite for symbolic communication does not necessarily imply that the ability to re-present only emerges when symbolic communication begins. One alternative course of psychohistorical evolution could be that the ability to form re-presentations is much older, but initially restricted to the system's "rest periods," that is, to times when no online control of behavior is required. In such a scenario, the ability to keep perception and re-presentation separate (= dual representation) would become indispensable only at the onset of symbolic communication. Even more far-reaching is the idea that a completely developed dual representation architecture already exists before the onset of symbolic communication. In this case, one would have to make other factors that have nothing to do with communication responsible for the formation of this ability. As soon as symbolic communication appears, the existing dual representation architecture can be used to interpret messages as external communication acts, and, secondarily, it can also be used to interpret thoughts as internal communication acts arising in the self. Hence, the common features that are critical for all conceivable variants of this scenario are: (a) the use of the dual representation architecture to interpret acts of communication (perception of the communication act with simultaneous re-presentation of the communicated content) and (b) the extension of this interpretation framework to cover internally induced re-presentation.

transition from one solution to the other and the mentalities associated with them are the subject of Julian Jaynes's speculative theory of consciousness. He even considers that this transfer occurred during historical times: between the Iliad and the Odyssey. In the Iliad, according to Jaynes, the frame of mind of the protagonists is still structured in a way that does not perceive thoughts, feelings, and intentions as products of a personal self, but as the dictates of supernatural voices. Things have changed in the Odyssey: Odysseus possesses a self, and it is this self that thinks and acts. Jaynes maintains that the modern consciousness of Odysseus could emerge only after the self had taken over the position of the gods (Jaynes, 1976; see also Snell, 1975).

Moreover, it is obvious why the *political* implications of the two solutions differ so greatly: Societies whose members attribute their thoughts to the voices of mortal or immortal authorities produce castes of priests or nobles that claim to be the natural authorities or their authentic interpreters and use this to derive legitimization for their exercise of power. It is only when the self takes the place of the gods that such castes become obsolete, and authoritarian constructions are replaced by other political constructions that base the legitimacy for their actions on the majority will of a large number of subjects who are perceived to be autonomous.

Finally, an important *psychological* difference is that the development of a self-concept establishes the precondition for individuals to become capable of perceiving themselves as persons with a coherent biography. Once established, the self becomes involved in every re-presentation and representation as an implicit personal source, and just as the same body is always present in every perceptual situation, it is the same mental self that remains identical across time and place.

3.3. Application to action: Goals and their sources

We now move from cognition to action, and thus from cognitive to dynamic re-presentation. What is the role of action plans and action goals in the process of forming a self? Their involvement in this process leads to an extended self that functions not only as a *representational center for thoughts* but also as a *decision making center for actions.*

To appreciate the importance of this development, let us return to our fictitious animal that still lacks dual representation. As we have seen, it controls its behavior through a continuous evaluation of incoming information with respect to its needs. As long as dual representation has not formed, needs cannot be represented alongside the contents of perception, but only *within* them. Perception is, so to speak, rendered dynamic by representing action incentives that conform to the need *within the current perceptual situation*. In other words, the contents of perception are equipped with attributes that specify their appropriateness in each case for satisfying current needs. Foodstuffs are enticing; partners, attractive; rivals, offensive; and one's predators, threatening. The organism is exposed to forces that are inherent in the perceptual world, and the animal's behavior is the outcome of the interplay of these forces (e.g., Lorenz, 1937, 1943; McDougall, 1908; Tinbergen, 1951).

Dual representation. However, the situation changes as soon as dual representation has formed. The interplay of forces is now accompanied by the interplay of thoughts—and, with this, the ability to form and maintain goals. The perception of current events is now joined by an independent re-presentation *of goals*, that is, of potential future events that are desired and striven toward. The ability to implement goals opens up completely new possibilities for controlling actions in line with needs, because it detaches the re-presentation of goals completely from the

representation of the current perceptual situation—making it possible to maintain explicit goal settings independently from the current perceptual situation. Actions can then be selected and configured so as to bring perceived actual states closer to explicitly re-presented desired states.

Attribution of goals to persons. When goals control actions, the question regarding where the goals come from is no longer just an attribution problem that is of only psychological interest, but becomes an issue of enormous political significance. Naturally, an attribution problem will also arise here only when action goals have been formed that are not given by other actors in acts of communication. In the latter case, there is no need for any additional attribution, because the person who is the source can be perceived in the current situation. The attribution problem will arise only when goal re-presentations do not come from outside, but emerge within the system itself.

In principle, the solutions are the same as before—but with the one difference that there is a much stronger political interest in society's regulation of this attribution process: In the long run, it concerns not just thoughts but real actions. One solution locates the source of action goals in the will of invisible personal authorities, that is, in external authorities who, somehow or other, *read into* the actor what he or she has to do and demand *obedience* by dint of their authority. The other solution, in contrast, locates the source of action goals in the self. This constructs an internal personal authority who reaches autonomous *decisions* on what he or she wants to do. Obedience is replaced by autonomy.

Therefore, cultures that place the self in the position previously reserved for gods or kings will bring forth autonomous agents—agents who understand their thoughts and actions, as it were, proceeding outward from within. As a cognitive and dynamic center, the self is simultaneously the source and the integration center for the person's psychological and physical activity.[5] Of course, this aristocratic role has its price: The mental self is now responsible for the thoughts and acts that proceed from it and can also be made liable for them.

3.4. Summary

This completes our psychohistorical scenario. As can be seen, it makes a lot of demands on our theoretical intuitions. First, it expects us to say goodbye to the idea that our selves are naturally given organs of our minds that form the foundation on which the entirety of our experience is based. Second, it expects us to believe that the self is nothing more than a specific mental content, based on knowledge structures formed in learning processes and shaped in social interaction—basically no different from the mental representations of objects and events in the external world (cf., e.g., Kihlstrom & Klein, 1994).[6] Third, it expects us to acknowledge that subjectivity,

[5] Just like the cognitive self, the dynamic self is generally also present only implicitly in the psychological acts concerned: A person who is planning acts or reaching action decisions is generally just as much "on the ball" as when he or she is watching something or thinking about something. Whereas the mental self is not involved explicitly in these acts, it is present implicitly. Then the "ball" that the planning person is "on" does not exist for its own sake, but as a ball that is attended *by the person him- or herself*—and it is precisely in this sense that it is conscious.

[6] The only feature through which the mental representation of the self stands out from mental representations of events in the world is its meta-representational status: The self is represented as the *source* of other mental contents and those contents are, in turn, always represented with reference to that source.

or selfhood, is injected outward-in (rather than projected inward-out, as most theories submit), that is, that it is first perceived in external sources before it then becomes established as an internal source as well.

4. Discourses on subjectivity

The conclusion so far is that the self is an invention for solving an attribution problem. Initially, it is set up as a source of internally induced re-presentation. Once this has occurred, its implicit presence in mental acts forms both the functional and the foundational basis for the conscious character of mental contents.

However, this invention should certainly not be viewed as a heroic achievement by single individuals. Rather, selves are construed and maintained in concrete social exchange—in discourses on subjectivity and consciousness held within a culturally standardized interpretation framework that controls the socialization of individuals and attributes them with a self-morphic organization of their mental structures. I shall conclude by discussing three kinds of such subjectivity discourse: discourses of attribution, reflection, and demarcation.

4.1. Attribution discourses

Consider first attribution discourses in daily life. The most elementary communication mechanisms are based on direct face-to-face interactions in the *microsocial field* and are not even necessarily bound to a verbal communication. When all actors in a social grouping organize the way they deal with each other on the basis of mutual attribution of self-morphic organization, each one of them—and also every new actor who appears on the scene—is confronted with a situation in which a self-morphic role is made available by the activities of others. In a situation like this, the external attribution of this role may finally generate corresponding internal attributions in the actor him/herself, and he/she will eventually adopt the attributed self-role as his or her own.[7]

More complex communication mechanisms are based on verbally bound attribution discourses in the *macrosocial field*. The first one to mention is the discourse on psychological common sense, in other words, the everyday psychological constructs that cultures or speech communities use to explain the actions of their members.[8] For instance, Western folk psychology operates on the basis of a theory of human personality with an explicit self at its core that remains identical throughout life and serves as the organizational nucleus for all cognitions and actions. Equally relevant are the discourses of morality and law: They derive personal responsibility for individuals' behavior from viewing their selves as autonomous sources of action decisions. Such discourses focus far more on the dynamic foundations of the self than its cognitive foundations—which is none too surprising in light of their enormous significance for the social control of action.

[7] The development of subjectivity and an understanding of subjectivity in early childhood has been studied in detail over the last 20 years. See, e.g., Perner (1991), Rochat (1995, 1999), and Wellman (1990).

[8] See, e.g., Heider (1958) and Prinz (1997). In recent years, the Cambridge philosopher Martin Kusch has developed a theoretical framework that emphasizes the role of folk-psychology notions as social institutions, created and maintained in permanent interaction and discourse (Kusch, 1997, 1999).

Attribution discourses permeate our daily lives to an extent that we are scarcely able to grasp. Every day and every hour, major parts of our mental life are given over to our thoughts about actions—about our own actions and those of others that we are explaining and evaluating. It is also these considerations that determine a large part of our social exchange—an exchange that, for many persons, consists in talking about who did what when and why, and how these actions should be evaluated. And, last not least, we have been raised and we live in a vast world of fictitious stories abounding with such issues—stories that we encounter in books, movies, and TV programs with which we fill out the times, in which we temporarily put aside the real stories we are otherwise involved in.[9] Moreover, it is also such stories that we tell our children in order to explain how thinking and doing relate to each other; and it is such stories that they absorb so avidly in all cultures. What these offer to them (and to us) takes two forms: first, the explicit semantics inherent to the particular culture—its morals and customs, its values and norms, and even its myths and religious beliefs—and, second, and simultaneously, the implicit syntax of commonsense psychology that specifies how human actors function and, above all, how their actions relate to their thoughts and desires. In this sense, folk psychology acts as a social institution that regulates our thoughts and actions (Kusch, 1997, 1999).

Although narrative may be the basic form of attribution discourse, this does not rule out the existence of other forms. Moral discourse does not just tell stories but explains and evaluates actions according to general rules. This is even more true for legal discourse that aims to supplement narrative action explanations through reflective action evaluations.

4.2. Reflective discourses

In our cultural domain, the attribution discourses of daily life have long been supplemented and extended through intellectual and scientific discourses that reflect on the role and status of consciousness, and if we want to adopt a comprehensive perspective, we also have to view these reflective discourses as elements of the broader cultural discourse that constitutes and maintains subjectivity. Philosophy is particularly important here, and has always taken a leading role in this discourse. For several hundred years, it has been dominated by one doctrine that could be labeled—somewhat disparagingly—as a *fundamentalism of consciousness* (Prinz, 1996a). This was formulated particularly succinctly in Descartes' doctrine that only *cogito*, the self-experience of the human mind, could be viewed as the one indubitable foundation of human knowledge.

Descartes believed that accessing one's own facts of consciousness is a process that is structured much more simply than accessing the external world. When accessing one's own consciousness, the mind is, so to speak, at home—instead of being confronted with material entities from the outside world from which it fundamentally differs. Accordingly, first-person knowledge appears to be infallible. Such knowledge, it is held, arises from immediate awareness of real facts—and not from a process of representation for which the question of the relation between real and represented facts could be posed in any meaningful way. Hence, phenomena of consciousness are

[9] Interestingly, the role of narrative is not just emphasized (and often also celebrated) in postmodernist approaches in the social and cultural sciences, but has also long been the subject of a variety of different philosophical approaches. See, for example, Dennett (1992), Lübbe (1960), and Schapp (1959).

primary and fundamental for knowledge, whereas phenomena referring to the external world are secondary and derived. This doctrine forms, so to speak, the analytically reflected, ideological core of the modern discourse on subjectivity. It nurtures the idea that the mental self forms the logical origin of all knowledge about the world—a theory that is naturally completely incompatible with the present scenario.[10]

4.3. Demarcation discourses

Finally, I would like to give a brief sketch of some demarcation discourses—discourses that are partly attributive and partly reflective in nature and serve to ascertain the limits of subjectivity. They differentiate between that which belongs to the domain of normally developed subjectivity and that which lies beyond. For example, psychopathology has long known that attribution to persons and self-constitution can take other than normal paths in the individual case—paths that are viewed as *pathological*.

One example is the formation of delusional symptoms in psychotic disorders, particularly in schizophrenia. According to the cognitive theories of schizophrenia developed in the last decade (Daprati et al., 1997; Frith, 1992), these symptoms can be explained with the same basic pattern that Julian Jaynes uses in his theory to characterize the mental organization of the protagonists in the Iliad. Patients with delusions suffer from the fact that the standardized attribution schema that localizes the sources of thoughts in the self is not available to them. Therefore, they need to explain the origins of their thoughts, ideas, and desires in another way (see, e.g., Stephens & Graham, 2000). They attribute them to person sources that are present but invisible—such as relatives, physicians, famous persons, or extraterrestrials. Frequently, they also construct effects and mechanisms to explain how the thoughts proceeding from these sources are communicated, by, for example, voices or pictures transmitted over rays or wires, and nowadays frequently also over phones, radios, or computers.

Another example of self-constitution that does not comply with the norm is the multiple-personality syndrome. The term multiple personality is used when two or more independent personalities have formed within one person, and each lead their own lives. Even though some particularly spectacular case reports have proved to be exaggerated and, in part, even fakes (Confer & Ables, 1983), the occurrence of such split personalities is well-documented (e.g., Kluft, 1991; for historical records, see Greaves, 1993; Oesterreich, 1910). To some extent they form the opposite pole of delusional symptoms: in the latter, the problem is a lack of self or self-attribution; in the former, that several selves are available at the same time.

Further observations on disorders in the customary relation between actor and action are known from split-brain patients and from healthy persons under posthypnotic suggestion. It has been reported that patients who have undergone surgical division of the corpus callosum connecting the two brain hemispheres sometimes say or do things without being conscious of them,

[10] It should be noted in passing that this doctrine is now being questioned just as much as it has venerated. Recent developmental studies have made a major contribution to these doubts. It has been shown that at no point in their development do infants possess a better understanding of their own mental states than those of other persons. It often seems as if even the opposite is true, and that self-understanding *follows on* from an understanding of others—a finding that cannot be easily reconciled with the doctrine of direct and privileged access to one's own mental states (e.g., Gopnik, 1993).

and then justify and rationalize these acts retrospectively. Unconscious performance and conscious rationalization are each assigned to one of the two divided hemispheres (see, e.g., Gazzaniga & Smylie, 1984; Zaidel, 1990). Similar observations are also reported regularly for the performance of posthypnotic tasks: When participants carry out tasks that have been assigned under hypnosis, they often provide attendant rationalizations for their (at times, highly comical) actions. It seems as if the act precedes the desire here; as if patients and participants do not do what they want, but rather (also) want what they (in any case) do or have done (see Prinz, 1996b).

As bizarre as these syndromes seem against the background of our standard concept of subjectivity and personhood, they fit perfectly with the theoretical idea that mental selves are not naturally given but rather culturally constructed, and in fact set up in, attribution processes. The unity and consistency of the self are not a natural necessity but a cultural norm, and when individuals are exposed to unusual developmental and life conditions, they may well develop deviant attribution patterns. Whether these deviations are due to disturbances in *attribution* to *persons* or to disturbances in *dual representation* cannot be decided here. Both biological *and* societal conditions are involved in the formation of the self, and when they take an unusual course, the causes could lie in both domains.

I shall finish with a twofold train of thought: This addresses a favorite question to be found at the center of the demarcation discourse—namely, whether and how far other living beings apart from *H. sapiens* possess self-morphic mental organization and consciousness, and whether it is also possible that there are human beings who do not possess this.

First, can a consciousness develop in animals if we attribute it to them? Would, for example, my dog Max develop a self-morphic organization if he interact exclusively with human beings who treated him like a conscious self? This notion conceals the question whether the social availability of attribution to persons is *sufficient* for the formation of a self-morphic mental organization. Following the psychohistorical sketch developed above, we have to answer this question negatively—at least as long as we do not wish to assume that the second necessary precondition has also formed in dogs: namely, the ability to maintain dual representations. Attribution to persons, as this example shows, is a *necessary* but not a *sufficient* condition for the formation of consciousness.

Second, can human beings become unconscious zombies when denied all interactions and discourses containing proposals for attribution to persons? For example, would it have been possible for the German foundling Kaspar Hauser to have been completely self-less and thus without consciousness? Our theory has to answer this question affirmatively, because it assumes that self-morphic organization and consciousness cannot emerge without socially mediated attributions.

Thus, it seems as though it may well be possible for human beings to live without consciousness, but not for animals to develop consciousness: Zombies may exist, but Bambi, Lassie, and Fury will always remain a charming illusion.

References

Brentano, F. (1924). *Psychologie vom empirischen Standpunkt (Bd. 1) [Psychology from an empirical perspective]*. Leipzig: Meiner (Original work published 1874).

Confer, W. N., & Ables, B. S. (1983). *Multiple personality: Etiology, diagnosis, and treatment*. New York: Human Sciences Press.

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Daléry, J., & Jeannerod, M. (1997). Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition, 65*, 71–86.

Dennett, D. C. (1990). The origin of selves. Report 14/1990 of the Research Group on Cognition and the Brain at the ZiF, University of Bielefeld, Germany.

Dennett, D. C. (1992). The self as the center of narrative gravity. In F. S. Kessel, P. M. Cole, & D. L. Johnson (Eds.), *Self and consciousness: Multiple perspectives* (pp. 103–115). Hillsdale, NJ: Erlbaum.

Donald, M. W. (2001). *A mind so rare: The evolution of human consciousness*. New York, London: W.W. Norton.

Edelman, G. M. (1989). *The remembered present: A biological theory of consciousness*. New York, NY: Basic Books.

Frith, C. (1992). *The cognitive neuropsychology of schizophrenia*. Hillsdale, NJ: Erlbaum.

Gazzaniga, M. S., & Smylie, C. S. (1984). What does language do for right hemisphere? In M. S. Gazzaniga (Ed.), *Handbook of cognitive neuroscience*. New York: Plenum Press.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *The Behavioral and Brain Sciences, 16*(1–15), 90–101 [Repr. in Goldman, A. (Ed.). (1993). *Readings in philosophy and cognitive science*. Cambridge, MA: MIT Press].

Greaves, G. B. (1993). A history of multiple-personality disorders. In R. P. Kluft & C. G. Fine (Eds.), *Clinical perspectives on multiple-personality disorder* (pp. 355–376). Washington, DC: American Psychiatric Press.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Jaynes, J. (1976). *The origin of consciousness in the breakdown of the bicameral mind*. Boston, MA: Houghton Mifflin.

Kihlstrom, J. F. (1997). Consciousness and me-ness. In J. D. Cohen & J. W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 451–468). Hillsdale, NJ: Lawrence Erlbaum.

Kihlstrom, J. F., & Klein, S. B. (1994). The self as a knowledge structure. In R. S. Wyer & T. K. Srull (Eds.), *Basic processes: Vol. 1. Handbook of social cognition* (2nd ed., pp. 154–208). Hillsdale, NJ: Erlbaum.

Kluft, R. P. (1991). Multiple-personality disorder. *American Psychiatric Press Annual Review of Psychiatry, 10*, 161–188.

Kusch, M. (1997). The sociophilosophy of folk psychology. *Study in the History and Philosophy of Science, 28*, 1–25.

Kusch, M. (1999). *Psychological knowledge: A social history and philosophy*. London: Routledge.

Lorenz, K. (1937). Über die Bildung des Instinktbegriffs. [On the formation of the instinct concept]. *Naturwissenschaften, 25*, 289–331.

Lorenz, K. (1943). Die angeborenen Formen möglicher Erfahrung. [The innate forms of potential experience]. *Zeitschrift für Tierpsychologie, 5*, 235–409.

Lübbe, H. (1960). Sprachspiele und Geschichten Neopositivismus und phänomenologieim Spätstadium. *Kant-studien, 52*, 220–243.

McDougall, W. (1908). *An introduction to social psychology*. London: Methuen.

Metzinger, T. (1993). Subjekt und Selbstmodell. Die Perspektive phänomenalen Bewußtsein vor dem Hintergrund einer naturalistischen Theorie mentaler Repräsentationen. [Subject and self-model: The perspective of phenomenological consciousness against the background of a naturalistic theory of mental representations]. Paderborn: mentis [out of print, 2nd ed., 1999].

Metzinger, T. (Ed.). (2000). *Neural correlates of consciousness: Empirical and conceptual questions*. Cambridge, MA: MIT Press.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.

Mithen, S. (1996). *The prehistory of mind. A search for the origins of art, religion, and science*. London: Thames and Hudson.

Oesterreich, K. T. (1910). *Die Phänomenologie des Ich in ihren Grundproblemen*. Leipzig: Johann Ambrosius Barth.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Prinz, W. (1996a). Bewusstsein und Ich-Konstitution. [Consciousness and the constitution of the self]. In G. Roth & W. Prinz (Eds.), *Kopf-Arbeit. Gehirnfunktionen und kognitive Leistungen* (pp. 451–467). Heidelberg: Spektrum.

Prinz, W. (1996b). Freiheit oder Wissenschaft? [Freedom or science?]. In K. Foppa & M. von Cranach (Eds.), *Freiheit des Entscheidens und Handelns* (pp. 86–103). Heidelberg: Roland Asanger.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*(2), 129–154.

Rochat, P. (1995). *The self in infancy: Theory and research*. Amsterdam: Elsevier.
Rochat, P. (Ed.). (1999). *Early social cognition: Understanding others in the first months of life*. Mahwah, NJ: Erlbaum.
Schapp, W. (1959). *Philosophie der Geschichten. [Philosophy of stories]*. Leer: Rautenberg.
Snell, B. (1975). *Die Entdeckung des Geistes. Studien zur Entstehung des europäischen Denkens bei den Griechen [Discovery of the mind. Studies on the emergence of European thinking in the Greeks]*. Göttingen: Vandenhoeck & Ruprecht.
Stephens, G. L., & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. Cambridge, MA: MIT Press.
Tinbergen, N. (1951). *The study of instinct*. London: Oxford University Press.
Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.
Whiten, A. (2000). Chimpanzee cognition and the question of mental re-representation. In D. Sperber (Ed.), *Meta-representations* (pp. 139–167). Oxford: Oxford University Press.
Zaidel, E. (1990). Language functions in the two hemispheres following complete cerebral commissurotomy and hemispherectomy. In F. Boller & J. Grafman (Eds.), *Hand book of neuropsychology* (Vol. 4, pp. 115–150). Amsterdam: Elsevier.