THE PARADOX OF SELF-CONSCIOUSNESS REVISITED

BY

PHILIP ROBBINS

Abstract: The so-called paradox of self-consciousness suggests that self-consciousness, understood as the capacity to think about oneself in a first-person way, cannot be explained. The author of the paradox contends that the only way to avert this result is by invoking the notion of nonconceptual first-person thought. This contention is rooted in adherence to the Linguistic Priority Principle, which dictates that pre- and nonlinguistic creatures lack concepts. I argue that the latter claim is dubious, and that the paradox of self-consciousness can be better disarmed by denying the dependence of first-person conceptual thought on first-person language.

The term ‘self-consciousness’, like the term ‘consciousness’, is used in a variety of senses. According to one usage, arguably the philosophers’ favorite, it refers to the capacity to detect and reflect upon one’s mental states (Van Gulick 1988). According to a second, self-consciousness is the capacity to think of oneself as one element among others in an objective spatio-temporal order (Evans 1982). A third usage has it that self-consciousness is the capacity to think of oneself as an entity with bodily and psychological properties which exhibit the causal structure characteristic of persons (Campbell 1994). According to a fourth, more abstract sense of the term, self-consciousness is the capacity to think first-person thoughts (Bermúdez 1998). Each of these types of self-consciousness is of interest in its own right, as are the relations among them. But it is self-consciousness in the lattermost sense – call it “core self-consciousness” – that will be my topic here.

Exactly what core self-consciousness comes to is controversial, but it can be safely presupposed to have at least two features. First, it is genuinely cognitive, insofar as it involves the ability to entertain a certain range of thoughts. Second, thoughts in this range are standardly expressed by
means of first-person linguistic utterances. These two remarks may not seem like much to go on, but they suffice to point toward a potential source of philosophical trouble – in particular, a paradox. My aim in this paper is to clarify what that trouble is, where it comes from, and how best to avoid it.

The post-expository discussion has both a critical and a constructive component. In the critical part, I argue against the claim that the problem in question – the so-called “paradox of self-consciousness” – has only one solution, namely, that favored by the author of the paradox (§§ 2–3). These negative remarks then serve to motivate, in the second, positive part of the paper, an alternative solution to the problem which is plausibly superior to the first (§§ 4–5). My main objective in this part is to show how core self-consciousness might be conceptual in character yet autonomous with respect to the language faculty. The interest of this demonstration, however, extends beyond the paradox of self-consciousness, insofar as it promises to shed new light on a bedrock component of our conceptual life.

1. Introducing the paradox

The paradox of self-consciousness suggests that core self-consciousness is in principle beyond the reach of psychological explanation. To see how this manifestly bizarre and unpalatable conclusion comes about, consider the following argument (PSC, 24):

(P1) The only way to explain core self-consciousness is by explaining the capacity to think first-person thoughts.
(P2) The only way to explain the capacity to think a certain type of thoughts is by explaining the capacity for the canonical linguistic expression of thoughts of that type.
(P3) First-person thoughts are canonically expressed by means of the first-person pronoun.
(P4) The capacity to think first-person thoughts is an essential ingredient of competence with the first-person pronoun.
(P5) If P1–P4 are true, then core self-consciousness cannot be explained in a noncircular way.
(P6) Noncircularity is a necessary condition on genuine explanation.

(C) Core self-consciousness cannot be explained.

The argument is plainly valid, and each of the six premises is plausible on its face – hence the paradox. At least one of the premises, however, calls for closer inspection.

The suspicious candidate is what Bermúdez dubs the
Thought-Language Principle. The only way to explain the capacity to think a given type of thoughts is by explaining the capacity for the canonical linguistic expression of thoughts of that type.

The Thought-Language Principle has had some distinguished adherents (see e.g. Dummett 1973). Nonetheless, the step in the paradoxical reasoning which it underwrites is far from secure. An obvious way to block it would be to argue that, at least with respect to certain types of thought, the capacity to think thoughts of those types is dissociable from the capacity to express those thoughts linguistically. For if such dissociability could be shown, that would effectively undermine the idea that the latter capacity must be explanatorily prior to the former.

As Bermúdez observes, however, skepticism about this strategy for solving the paradox can be induced on the basis of the following considerations (PSC, 41–42):

Conceptual Requirement Principle. The range of thoughts attributable to a subject is exhaustively determined by the range of concepts in that subject’s possession.

Linguistic Priority Principle. Conceptual abilities are constitutively linked with linguistic abilities in such a way that conceptual abilities cannot be possessed by pre- or nonlinguistic subjects.5

These two further principles jointly entail the impossibility of thought without language. So if the Thought-Language Principle is to be given up on the grounds suggested, then at least one of them will have to be given up as well. Trouble is, it’s not clear which – if either – we can get by without.

On this score, Bermúdez makes two points: first, that the paradox of self-consciousness can be solved by rejecting the Conceptual Requirement Principle (CRP); second, that this is the only method of resolution available, given the options indicated above (PSC, 272). Apropos of the first point, it is argued that there exists a type of first-person thoughts which are not conceptually articulated, hence available to subjects who lack mastery of the first-person pronoun. The second point rests upon an argument which purports to show that the Linguistic Priority Principle (LPP) is non-negotiable – leaving us with no choice but to abandon CRP.

In the discussion ahead, I will remain agnostic on the first point. That is, I wish to leave open the possibility of a nonconceptualist solution to the paradox. What I will dispute is the purported necessity of that solution. Indeed, I hope to show that core self-consciousness can be better explained within a conceptualist framework.

© 2002 University of Southern California and Blackwell Publishers Ltd.
2. An a priori argument for priority

The Linguistic Priority Principle (LPP) makes a strong and unobvious claim about the nature of conceptual representation. Bermúdez defends the principle on a priori grounds by arguing that concept-involving modes of explanation do not apply to the behavior of creatures without language, since languageless creatures lack the minimum epistemic credentials for concept possession (PSC, 69–71). Here is how the argument goes:

(P1) Explanations of behavior which advert to a subject’s conceptual resources are acceptable only when the subject can employ the relevant concepts in rational judgment.

(P2) The capacity for rational judgment requires both the capacity to form justified beliefs and the capacity to justify (some of) those beliefs.

(P3) The capacity to justify one’s beliefs requires linguistic competence.

(C) Concept-involving modes of explanation do not apply to pre- or nonlinguistic creatures.

Though the reasoning above is valid, its soundness can be questioned at various points. I’ll indicate three of them.

First, it is important to distinguish between the ability to justify a belief – that is, to have beliefs with justificatory content – and the ability to communicate those beliefs linguistically. Failing to observe this distinction begs the question as to whether concepts are dissociable from language. Yet this is just what Bermúdez does when he claims that:

[providing justifications is a matter of identifying and articulating the reasons for a given classification, inference, or judgement. It is because prelinguistic creatures are in principle incapable of providing such justifications that the priority principle is true. (PSC, 71)]

The core claim here is that “identifying and articulating” the justificatory basis of a judgment is an essentially linguistic activity. But terms like ‘identify’ and ‘articulate’ are equivocal: they admit what I shall call, following Sperber and Wilson 1995, both cognitive and communicative readings. Cognitively identifying or articulating a justification is something done in thought, the communicative counterpart of which is something done in language. Of course, it might be argued that cognitive justification is parasitic on justification of the communicative kind; but all we get on this point from Bermúdez is assertion.

A second, more pressing, point is this. Even if we grant that the ability to justify (hence, form beliefs about) one’s beliefs requires language, it is hardly obvious that creatures lacking metacognitive abilities of this sort cannot be true believers. To see why, consider reliabilism in epistemology...
(Goldman 1979). According to reliabilists, whether or not a belief is justified may depend solely on its causal ancestry, that is, whether it was produced by a reliable (i.e., truth-tracking) psychological process. There is no requirement here that the believer must be able to identify that ancestry in general. Why insist on this ability in so much as a single case? At this point Bermúdez simply appeals to epistemological intuition (PSC, 71). But intuitions of this sort are not bedrock. On the contrary: they are notoriously unstable. And the particular intuition at issue is shaky at best. What’s needed is further argument, and none is on offer.⁶

Third, a methodological objection. It seems clear that the question as to whether concepts are possible without language is a broadly empirical one. So we should expect it to be settled largely on a posteriori grounds, not from the armchair. A priori reasoning may play a role here, but probably no more than a peripheral one. That Bermúdez lays such an emphasis on intuitions about the nature of justification – a decidedly elusive topic – is somewhat surprising, given his repeated insistence on the naturalistic slant of his inquiry.⁷ After all, the very idea of naturalism in philosophy militates against this way of proceeding (Chomsky 1995).

Given these doubts about Bermúdez’s brief for LPP, it would be unwise for us to endorse the principle on that basis – especially in light of empirical evidence against it, to which I now turn.

3. Against priority

The Linguistic Priority Principle looks even less inevitable when viewed against the background of recent work in developmental and comparative psychology. Though a thorough survey of the evidence is beyond the scope of this paper, I’ll briefly sample some research in two areas – object cognition and mirror self-recognition – which does not sit well with the principle.

3.1 OBJECT COGNITION

Cognitive psychologists have long been interested in the development of our abilities to think and reason about physical objects. Recent evidence, collected from experiments which employ the so-called “preferential looking” paradigm, suggests that these abilities are in place as early as 4 to 5 months of age – well before language acquisition sets in. Here is how the paradigm works. First, infants are habituated to an initial event. This involves their being shown the event repeatedly until their looking times decline to a pre-set level (e.g., half what it was at the outset). Then the infants are dishabituated to the event: they are shown a pair of outcomes in succession – one of which matches adult expectations and another
which does not – and their looking times for each display recorded. If infants look longer at the unexpected outcome, this supports the attribution of a cognitive capacity akin to that found in adults.

For example, in the “drawbridge” experiment reported in Baillargeon et al. 1985, infants aged 5 months were presented with a screen facing them, then habituated to the screen’s being rotated vertically 180 degrees back and forth, toward and away from the subject. At that point, a box was displayed and placed behind the face-down screen. Rotation of the screen away from the infant was then continued, either with the screen stopping short of 180 degrees (expected outcome) or with the screen going all the way down (unexpected outcome). Trial after trial, infants attended significantly longer to the full-rotation event. On the basis of results like these it has been argued that pre-verbal infants possess the concept of a physical object, where this concept is sustained by knowledge of basic physical principles, such as the principle that two objects cannot occupy the same place at the same time. According to this robustly conceptualist view, “the infant’s mechanism for apprehending objects is a mechanism of thought: an initial theory of the physical world whose . . . principles jointly define an initial object concept” (Spelke 1988, p. 181).8

Taken in isolation, however, the original drawbridge studies do not tell very strongly against priority. The main reason for this is the recent development of plausible explanations of the looking-time data which appeal solely to the operation of subconceptual mechanisms – specifically, mechanisms of visuo-spatial attention (Leslie et al. 1998, Scholl and Leslie 1999). Nonetheless there is a solid case to be made that nonconceptualist theorizing about object cognition has serious limitations. Four points are especially salient here.

First, unlike mechanisms in the perceptual system, it is a trademark of conceptual mechanisms that their operation is directly sensitive to the flow of information elsewhere in the cognitive system. Conceptual mechanisms, it is said, are typically “cognitively penetrable”, whereas subconceptual mechanisms are not (Pylyshyn 1984). And some aspects of object cognition in infants exhibit this feature of cognitive penetrability. For example, a later drawbridge experiment showed that infants do not look longer at the full-rotation outcome when allowed to discover beforehand, through haptic manipulation, that the box placed behind the screen is compressible (Baillargeon 1987). This suggests that infants’ tracking of contact-mechanical properties of objects is at least partly subserved by a conceptual mechanism.

Second, as noted above, one thing we wish to explain is why infants’ looking behavior accords with the principle that two objects cannot be in the same place at the same time. But it is difficult to do this at the subconceptual level because perceptual and attentional systems do not
always operate in accord with this constraint. This is evident from studies of the “Pulfrich double pendulum illusion”, in which viewers of a specially prepared experimental apparatus seem to see two rigid rods passing through one another (Leslie 1988).

Third, it appears that object cognition in infants accords with the same constraints in the haptic modality as it does in the visual one (Spelke 1988, Spelke and Van de Walle 1993). For example, studies done with infants aged four and a half months indicated that these subjects expect the behavior of haptically presented objects to conform to the principle of contact (two objects move together only if they touch) and the principle of cohesion (two surfaces lie on a single object just in case they are connected). This sort of cross-modal uniformity is *prima facie* (albeit weak) evidence that object cognition is subserved by a single, amodal mechanism which processes information derived from a variety of perceptual systems.9

Fourth, advocates of nonconceptualist accounts of object cognition concede that the explanatory range of this approach appears to be relatively narrow compared with that of the standard conceptualist alternative (Scholl and Leslie 1999). In particular, though nonconceptualist accounts do a respectable job of explaining (or explaining away) infants’ grasp of *spatiotemporal* principles, such as the principle of continuity (every moving object traces exactly one spatiotemporally connected path), it’s doubtful whether they can do the same with *mechanical* principles, such as the principle of contact. Almost everyone agrees that very young infants are able to track non-spatiotemporal, viz. contact-mechanical, properties of objects. But no one has a clue as to how to explain this in terms of the operation of subconceptual mechanisms.

On the basis of the foregoing considerations it seems fair to say that our best current understanding of object cognition in pre-verbal infants does not accord well with the Linguistic Priority Principle. One might still insist, however, that the principle does hold in the case of certain concepts. In particular, one might insist that possession of the self concept requires competence with the first-person pronoun – thereby blocking the possibility of a conceptualist solution to the paradox of self-consciousness. But here too, as we’re about to see, the empirical record suggests otherwise.

3.2 MIRROR SELF-RECOGNITION

Evidence for the dissociability of concepts from language in the first-person case comes from studies of the ability to recognize oneself in a mirror. In the “mark test” paradigm first documented in Gallup 1970, a subject whose forehead has been surreptitiously marked with dye is situated in front of a mirror and her actions observed. If the subject reaches
up to examine the mark or engages in other self-exploratory behaviors using the reflected image as a guide, it is likely that self-recognition has occurred. For it is natural to explain these behaviors by saying that the subject has identified herself self-consciously as the individual reflected in the mirror.

The relevant empirical facts are these:

- Developmentally normal humans recognize themselves in mirrors as early as 15 months of age, but they do not begin to use personal pronouns before 24 months (Lewis and Brooks-Gunn 1979);
- At least two non-language-using species, viz. chimpanzees and orangutans, recognize themselves in mirrors (Povinelli and Prince 1998).

As in the case of the object cognition data, the default interpretation of this data is a conceptualist one. One cannot recognize oneself in a mirror without identifying oneself as a certain individual, and one cannot identify oneself as a certain individual without coming to believe oneself to be that individual. Since belief states are, by consensus, conceptually articulated, it follows that mirror self-recognition (MSR) is a concept-involving feat. To forestall this conclusion, advocates of linguistic priority need to supply a non-conceptualist account of MSR. But how such an account would go is anyone’s guess. It’s probably no accident that in the thirty years since MSR studies were first conducted, no account of this kind has been advanced in the empirical literature.\(^\text{10}\)

This sort of worry ought to be especially pressing for a pro-priority theorist like Bermúdez, given certain of his other theoretical commitments. One reason for this emerges from his discussion of a family of neuropsychological disorders called “visual agnosias”, and in particular from his remarks about the taxonomic divide between “apperceptive” and “associative” agnosias. Bermúdez glosses this distinction as follows:

In very broad terms, the distinction is between deficits in object perception (the apperceptive agnosias) and object recognition (the associative agnosias). Apperceptive agnosias are due to impairments in visual perception that, although at a higher level than, for example, visual-field deficits, nonetheless appear to be perceptual rather than recognitional. Objects are not properly perceived, and hence are not recognized . . . in the associative agnosias, on the other hand, objects do seem to be properly perceived but are nonetheless not recognized. Such deficits appear to be more cognitive than perceptual. (PSC, 79–80)

He then goes on to compare the visual world of the pre-verbal infant with that of an “idealized global associative agnostic – a patient whose representations of the world are devoid of semantic features,” adding that the components of the visual system which are damaged in associative
agnosics are “best viewed as generating representations of the world at the conceptual level” (PSC, 81). Finally, in a tacit allusion to LPP, he notes that associative agnosias are correlated with lesions in language-specialized areas of the brain.

These remarks make it clear that, in Bermúdez’s view, object recognition is a concept-involving capacity. But the capacity to recognize oneself, in a mirror or otherwise, is surely just a special case of that more general capacity. Impaired face recognition, or prosopagnosia, is standardly classified as an associative agnosia, and in severe cases, a prosopagnosic will regularly mistake her reflection in a mirror for that of a stranger (Pallis 1955). So if Bermúdez is right to insist on the essentially conceptual character of object recognition – which seems entirely plausible – then an adequate nonconceptualist explanation of MSR is probably not in the cards.

4. A conceptualist way out

If what I’ve said so far is on the right track, then we needn’t suppose that the Linguistic Priority Principle is true; so we needn’t suppose that the only way to escape the paradox of self-consciousness is by countenancing the possibility of nonconceptual first-person thought. Instead, all we need to do is tell an empirically plausible story about the self concept which does not tie possession of this concept to competence with the first-person pronoun. In this section I’ll make a start in that direction.

What I propose to do, more specifically, is to defend a pair of hypotheses about which features of the self concept distinguish it functionally from other items in a thinker’s conceptual repertoire. The first hypothesis singles out the self concept in terms of a privileged relation to input systems, specifically systems of internal monitoring (§ 4.1). The second assigns to the self concept a complementary role on the output side, in decision-making and the control of behavior (§ 4.2). On the picture that emerges, self-consciousness is depicted as a capacity sustained by relatively primitive mechanisms of awareness and action, mechanisms which are plausibly independent with respect to those responsible for the production and understanding of language. This picture has obvious importance beyond the paradox of self-consciousness, given the empirical evidence of conceptually articulated self-consciousness in pre- and nonlinguistic creatures.

4.1 self-consciousness and internal monitoring

To begin with, I want to elaborate a suggestion about how to think of conceptual representation in general. The idea is this. Each concept points to a file containing information, or misinformation, about its
referent. In the case of individual concepts, which refer to individuals (rather than properties or kinds), each concept is linked to an object file. An object file is a file which serves as repository for information about a particular object. It is essential to object files that the information they are liable to contain is unrestricted as to its source in the information-processing system of the host: one item may derive from perception, another from inference, another from memory, and so on. This informational generality is a hallmark of the object files involved in conceptual representation.

However, not all object files are alike. Just as we can distinguish among different sources of information, we can distinguish among object files in terms of their relations to those sources. One way to do this is via the notion of specialization, which can be glossed as follows:

(SS) Given an information source S and an object file F, F is specialized for S, or S-specialized, just in case all information derived from S is routed directly to F.

(S) An object file F is specialized just in case there is some information source S such that F is specialized for S; otherwise F is generic.

Note here that the fact that F is specialized for S does not mean that the only information in F is from S, since that would conflict with the idea that object files are informationally general. Nor does it mean that the only object files containing information from S are those which are S-specialized. The point is rather that S-specialized files are guaranteed to contain information from S, whereas non-S-specialized files may or may not contain this information. What underwrites such guarantees is simply the path of information flow. If a file is specialized, then there must be some source of information whose output flows directly into that file, with no mediate processing by the conceptual system; if there is no source which fits this description, the file is generic.

It seems intuitively clear that the object files associated with most individual concepts are generic, rather than specialized. Consider, for example, concepts encoded by proper names, like SUPERMAN and ACAPULCO. The object files associated with such concepts are not a plausible first stop for information en route from any particular source. Nonetheless, the distinction between specialized and generic object files is a non-empty one, since every self-conscious agent possesses at least one concept whose associated object file is of the specialized variety.

The concept I have in mind here is, of course, the self concept. Here is a first pass at the proposal:

Specialization Hypothesis (preliminary version). The self concept of an agent is linked to a specialized object file on the agent.

© 2002 University of Southern California and Blackwell Publishers Ltd.
If a concept is associated with a specialized object file, there must be some source or sources of information which are “dedicated” to that file. Naturally, in the case of the self concept, where the associated file is a file on the concept’s host, we should expect these to be sources of self-specifying information, viz. sources of information about the agent. Though there aren’t many candidates for this role, one possibility stands out. To explain what this is, I need to introduce a second distinction, familiar from information-processing accounts of perception.

The job of the perceptual system is to pick up information about the world (understood as including the agent), process it, and finally deliver it to the conceptual system, including the subsystem which is responsible for the fixation of belief. In short, perception provides the raw material for conception. Our perceptual system consists of various coordinate subsystems, each of which corresponds to a distinct perceptual modality. These subsystems can be sorted into two categories: exteroception and interoception. The distinction between exteroception and interoception can be understood roughly as follows. Exteroceptive information may be either about the environment or about the agent; interoceptive information, by contrast, is invariably about the agent, though some of it may also concern the environment. Thus, the domain of interoception is relatively internal to the agent, compared with that of exteroception.

There are five exteroceptive modalities, each with its own sensory channel: vision, audition, touch, taste, and smell. These five senses are what one ordinarily thinks of under the rubric of perceptual awareness, but the field of such awareness is arguably much greater. The category of interoception covers a wide range of other modalities. The range is so wide, in fact, that it will be convenient to divide this category into two subcategories: somatic and psychological. The division here is straightforward. Somatic interoception, or “proprioception”, yields information about the agent’s body; psychological interoception, or “introspection”, yields information about her (occurrent) mental states. The somatic modalities are an especially diverse lot, making available many different types of bodily information, including the following (cited verbatim from Bermúdez et al. 1995, p. 13):

- Information about pressure, temperature, and friction from receptors on the skin and beneath its surface;
- Information about the relative state of body segments from receptors in the joints, some sensitive to static position, some to dynamic information;
- Information about balance and posture from the vestibular systems in the inner ear and the head/trunk dispositional system and information from pressure on any parts of the body that might be in contact with a gravity-resisting surface;
• Information from skin-stretch about bodily disposition and volume;
• Information about effort and muscular fatigue from muscles;
• Information about general fatigue from cerebral systems sensitive to blood composition.

This is not a complete catalogue, but it should suffice to give the feel of the topic.

With this taxonomy of perceptual systems in place, we can resume our hypothesizing about the functional profile of the self concept. For present purposes the principal division in that taxonomy – namely, the distinction between exteroception and interoception – is the one that matters. As noted above, the deliverances of interoception, somatic or psychological, are invariably about the agent, whereas those of exteroception may or may not be. This fact suggests a natural way to strengthen our earlier proposal about the self concept:

Specialization Hypothesis (revised version). The self concept of an agent is linked to an object file on the agent which is specialized for interoception.

The plausibility of this strengthening can be defended as follows. Let’s assume that the weak version of the Specialization Hypothesis is correct, and an agent’s self concept is indeed associated with a specialized object file. The purpose of this file is to store information about the agent. From a design standpoint, this much is clear: If the self file is specialized, then whatever information source or sources for which it is specialized ought to be such that they reliably yield information about the agent, and not about some other object. Since interoceptual information always has this self-specifying character, sources like proprioception and introspection fit the bill nicely. Exteroceptual sources, like vision and hearing, are much less reliable in this respect. For example, if you and I are playing the piano together, I might mistake some of your fingers for mine when judging on the basis of visual cues; but using kinaesthetic feedback as my guide, I cannot make such errors. Thus, engineering considerations make interoception a likely candidate for the role of source for which the self file is specialized.

The bulk of evidence for the Specialization Hypothesis is, alas, merely anecdotal; yet there does appear to be considerable evidence of that sort. The cases I have in mind are simply those in which an agent consciously judges on the basis of interoceptual input that some state of affairs obtains. In such cases, where judgment is triggered by the flow of “inside information”, the thought entertained by the agent is invariably one with an explicit self-conceptual component.

Consider the typical profile of a conscious judgment anchored in proprioception (Perry 1990). You’re at a cocktail party, hovering near
the buffet, when you spot an errant canapé on the floor. Bending over to retrieve it, you are mortified to hear the sound of cloth tearing, from which you gather that your trousers are now, tragically, breached. Thanks to the activity of mechanisms which monitor skin temperature, you become aware of your face flushing hotly. Now, if you come to believe on this basis that someone’s face is flushing, then you cannot fail to judge in a self-conscious way that you are that person. Under these epistemic circumstances, in other words, there is no room for a non-self-conscious thought to the effect that face flushing has occurred.

The same goes for other judgments made on the basis of consciously processed proprioceptive information. If a subject detects hunger, thirst, or fatigue via somatosensory channels, and this detection prompts her to a judgment on the matter, then what she will invariably judge is that she herself is hungry, thirsty, or tired. This pattern of phenomena is neatly explained by the idea that proprioceptive information is routed directly to the object file associated with the subject’s self concept. An analogous pattern is exhibited by conscious judgments grounded in the introspection of occurrent mental states (thoughts, pains, itches, and such), which are likewise always self-conscious in character.13

4.2 SELF-CONSCIOUSNESS AND CONTROL

We’ve considered the distinctive role of the self concept in relation to the processing of perceptual information (broadly construed). But that is only part of the story. For the self concept also seems to play a distinctive role in the production of action.

I’ll begin with a bit of intuition pumping. Consider the following anecdote, inspired by Perry 1979:

Mort is riding on a department store escalator when he sees in the middle distance a man whose shirt is streaked with a fine white powder, like chalk dust. It occurs to Mort that the man is probably unaware of his condition, and that he’d want to correct it were he informed. So Mort decides to do him this courtesy. However, as he steps up to address the stranger, he realizes that he’s been scrutinizing his own reflection in a mirror – and straightaway begins to brush off the dust from himself.

Episodes of this sort are routine in daily life. What they suggest is that no functional account of self-consciousness can be complete without adverting to the special contribution of self-conscious attitudes to decision-making, and hence to the production of behavior.

When someone desires something, very often what they desire is an event (i.e., that such-and-such comes to pass). Actions are a species of events and as such constitute a natural class of objects of desire. Within
the class of actions that an agent desires there is a further division, namely, between actions performed by the agent and actions performed by others. For instance, I might want to drive a friend’s Ferrari, but I also might want her to sell it and donate the money to charity. The first desire targets something I do, the second targets something my friend does. Of course, the only actions over which an agent has any direct control are her own. So only desires of the first type have any sort of potential to issue directly in performance of the action desired.

What the anecdotal evidence suggests, however, is the need for a distinction within this first class of desires. For it seems that some desires for actions performed by an agent have action-generative potential in this sense, whereas others do not. If an agent’s desire for an action is to result directly in its performance, it is not enough that the action in question be hers. There’s a further requirement, namely, that she is thinking of herself self-consciously as the relevant actor. If the agent desires an action of hers in a non-self-conscious way, it is effectively the same as if she had desired someone else’s action. In other words, if A’s doing X is an action you’d like to see performed, you’re not going to do X yourself unless you’re thinking of A self-consciously, that is, as yourself.

Let’s see how this diagnosis applies to Mort’s case. When Mort sees the man in the mirror, he thinks:

(1) THAT GUY SHOULD BRUSH THE CHALK DUST OFF HIMSELF

Among Mort’s initial desires, then, is a desire for the man he sees, call him M, to dust himself off. Though Mort is that man, he doesn’t realize this; thus, in Mort’s notional world, the action in question is not one that he can perform.14 So he sets out to perform a different action, namely, informing M of the situation – this second action being intended to promote the performance of the first. Shortly thereafter, Mort realizes that he himself is M, and with this realization comes the self-conscious desire to dust himself off. Only then does the initially targeted behavior ensue.

The moral of this story can be schematized along the following lines. According to a standard conception of cognitive architecture, cognitive systems harbor a variety of functionally distinct representational states. These states include beliefs, which represent the agent’s assumptions (how she takes the world to be), and desires, which represent the agent’s goals (how she would like the world to be). Beliefs and desires are subject to processing in the decision-making system, whose task it is to determine which actions are to be undertaken at any given time. These determinations are then sent as output to the action control system, which is directly responsible for producing and guiding the agent’s behavior.

In the decision-making process, desires for action are converted into decisions to act. Desires of this sort correspond to representations of the form:
Des (A does X)

whereas decisions correspond to representations of the form:

Dec (Do X).

If the action desired is an action by the host, the conversion from (2) to (3) may be direct, that is, unmediated by further processing in the decision-making system. But such cases are constrained by the following:

Conversion Rule. A desire for A to do X can be directly converted into a decision to do X only if ‘A’ = ‘Self’.

That is, only desires for action of the form

Des (Self does X)

are directly convertible into (3). This rule ensures that all representations of action which are relayed to the host’s control system for execution represent actions of the host, as opposed to actions of other agents. So it makes good sense from an engineering point of view.

Using this framework, we can advance a further hypothesis about the functional profile of the self concept, to wit:

Conversion Hypothesis. The self concept of an agent is such that a desire for A to do X can be directly converted into a decision to do X only if the agent is mentally referring to A under this concept.

This hypothesis specifies, in a highly general form, the functional relation between self-consciousness and action. Such a claim seems plausible both on architectural grounds and in light of the available evidence.

4.3 CLARIFICATION AND CAVEAT

Taken together, the Specialization and Conversion hypotheses yield a functional account of the self concept which frees the latter from any dependence on language mechanisms. Such an account points to a way out of the paradox of self-consciousness which does not compel us to abandon the familiar picture of thoughts as conceptually articulated representations. What it does compel us to do, however, is to regard self-consciousness as a capacity which can be abstracted away from more complex capacities which, in addition to implicating the self concept, also implicate a particular range of other concepts, such as concepts of physical objecthood, causality, or agency.
On the present account, a subject turns out to be self-conscious just in case she possesses a concept with the functional profile sketched above and can use that concept in thinking about herself, whatever other concepts she has at her disposal. The story about self-consciousness on offer, therefore, is conceptualist only in a minimal sense; but that in itself should be no cause for alarm. On the contrary, it seems only appropriate to the nature of the case.15

5. Closing reflections

Let me pause to take stock before pressing on a bit further. Recall that, according to the author of the paradox of self-consciousness, the only feasible way to resolve the paradox is to reject the Conceptual Requirement Principle—in particular, by positing a type of first-person thoughts which are not composed of concepts. The main reason for this insistence lies with a commitment to the Linguistic Priority Principle, which precludes the possibility of a subject’s possessing the self concept without being able to employ the first-person pronoun. In reply I’ve argued that the paradox can also be resolved in a manner consistent with the idea that first-person thoughts are ipso facto conceptually articulated. If so, the claim that jettisoning CRP is the only way to escape the paradox cannot be sustained.

Now I want to suggest something stronger: namely, that given the choice between an escape strategy which gives up CRP and retains LPP (like Bermúdez’s), and one which retains CRP and gives up LPP (like mine), we do better to opt for the latter. There are two reasons for this.

First, there are the costs of investing in the idea that concept possession essentially depends upon linguistic ability, both in general and in the first-person case in particular. Despite the philosophical pedigree of this idea, its status in contemporary naturalist circles is dubious at best—given both the dearth of naturalistically kosher arguments in its favor and the weight of empirical evidence against it. An advantage of the pro-CRP strategy sketched above is that it enables us to avoid these costs.16

Second, both strategies suppose that some thoughts are built up from concepts; the question is whether all thoughts are like this. Pro-CRP strategists say yes, whereas anti-CRP strategists say no. It’s important to note the ontological implications of rejecting this principle. For in saying that some thoughts are nonconceptual, one commits oneself not just to the idea that there are nonconceptual mental states, an idea taken for granted by most theorists of perception and attention.17 Nor is it just the idea that there are more thoughts than previously supposed. Rather, in positing the existence of nonconceptual thoughts, one is committed thereby
to the idea that there are more kinds or types of thought. This becomes clear from the fact that conceptual and nonconceptual thoughts have fundamentally different properties. For instance, it’s generally assumed that conceptual thought is productive, in the sense that there is no finite upper bound on the number of distinct thoughts a thinker can potentially host. Nonconceptual thought is distinguished by the fact that it lacks this property (PSC, pp. 92–93). This is good prima facie evidence of a taxonomic or typological distinction between conceptual thought, on the one hand, and nonconceptual thought, on the other. But once such a distinction is admitted, the anti-CRP position can be seen to suffer on methodological grounds. For the principle of parsimony, otherwise known as Ockham’s Razor, dictates that types of entity are not to be multiplied beyond explanatory necessity. Ceteris paribus, the theory with the leaner ontology is to be preferred. And the leaner theory in this case is the one which respects CRP.

To sum up: Pending evidence of psychological phenomena which cannot be explained save by adverting to nonconceptual thoughts, or a sturdier argument for priority, the Conceptual Requirement Principle should be preserved. In the meantime we have no good reason to believe in nonconceptual first-person thought, and good reason not to.18

Department of Philosophy
Washington University in St. Louis

NOTES

1 Namely, the eponymous paradox of Bermúdez 1998. In subsequent references, this work will be cited as PSC.
2 Or functionally analogous linguistic device. So-called pro-drop languages, such as Spanish and Italian, allow for subject pronouns to be unexpressed. In such languages the verb inflection tends to be rich; even so, personal pronouns exist and are used, both nominatively and otherwise.
3 The root idea is this: If a speaker successfully refers to herself by uttering a token of the first-person pronoun, then she cannot fail to know in a first-person way that she is the referent of that token (Anscombe 1975). Competence with the first-person pronoun, that is, requires not just that the speaker be able to recognize herself as the referent of her tokens of that expression, but that she be able to do so self-consciously.
4 The term ‘paradox’ does not seem strictly appropriate here. A paradox is an argument whose soundness cannot be denied without doing real violence to intuition, and the argument under consideration does not meet that condition. A good example of an argument which does is the sorites:

i. Someone with 0 hairs is bald.
ii. If someone with \( n \) hairs is bald, then so is someone with \( n+1 \) hairs.
   Therefore,
iii. Everyone is bald.
Measured by the soritical standard, Bermúdez’s use of the term ‘paradox’ in the present setting seems like false advertising (‘reductio’ would be more apt). I adopt it here mostly for ease of exposition.

5 This corresponds to what Bermúdez calls simply the ‘Priority Principle’. I’ve changed the label for mnemonic purposes.

6 What’s more, recent work in naturalized epistemology points to the opposite conclusion. For example, Kornblith (forthcoming) argues at length against the denial of epistemic capacities to creatures lacking the capacity for introspective reflection or other forms of metacognition.

7 For programmatic statements to this effect, see PSC, xii and pp. 47–48.

8 Theoretical knowledge of this sort, like a speaker’s knowledge of syntax, is thought to be neither explicit nor consciously accessible, and may well be innate (Spelke and Van de Walle 1993).

9 The evidence is also compatible with the idea that the visual and haptic modalities incorporate separate mechanisms each of which operates in accord with these principles. But the single-mechanism hypothesis is at least marginally preferable on grounds of architectural simplicity.

10 It may also be no accident that Bermúdez never addresses, or even mentions, mirror self-recognition – despite ubiquitous discussion of the topic in the psychological literature. Even a cursory glance at this literature reveals how remarkable an omission this is, especially in philosophical work as avowedly naturalistic as Bermúdez’s. Almost every developmentalist writing about self-consciousness, it seems, has something to say about MSR; see, for instance, Butterworth 1990, Kagan 1998, Meltzoff 1990, and Neisser 1993, among many others. The situation in primatology is the same; see references in Povinelli and Prince 1998.

11 Here and elsewhere I follow the convention of indicating names of concepts with small capital letters.

12 Note that, though introspective capacities may or may not be present in pre- or non-linguistic creatures, proprioceptive capacities surely are.

13 Much the same point is made in a classic work by William James: “Every thought tends to be part of a personal consciousness . . . It seems as if the elementary psychic fact were not thought or this thought or that thought, but my thought, every thought being owned . . . The universal conscious fact is not ‘feelings and thoughts exist’ but ‘I think’ and ‘I feel’” (1981, p. 221). James’s concern here, I take it, is with what might be called “introspective” self-consciousness – the capacity to reflect on one’s mental states, or to “read one’s own mind” (Nichols and Stich, forthcoming).

14 This is, of course, an artifact of the reflexive character of the action: though someone other than M could brush him off, only M can do this and brush himself off in a single act.

15 The specialization-and-conversion view of self-consciousness is not without precedents in the philosophical literature. In fact, in various writings it has been suggested that the self concept is constitutively tied to perception (Evans 1982, Recanati 1993), or to action (Perry 1979, Millikan 1990), or to both perception and action (Perry 1990, Rey 1997). Entries in this third category – what might be called “bilateral” accounts – are of particular interest, since they structurally resemble the account on offer here. For discussion of what distinguishes the specialization-and-conversion account from competing bilateral accounts, see Robbins 2000, pp. 167–71.

16 An alternative cost-saving strategy for resolving the paradox of self-consciousness would be to drop both LPP and CRP. A hybrid strategy of this sort is worth considering, but for reasons of space I can’t discuss it here.
442 PACIFIC PHILOSOPHICAL QUARTERLY

17 That much should be clear from the fact that, if we were to read the Thought-Language Principle as a claim about the explanatory priority of language with respect to, say, vision or audition, the principle would have zero initial plausibility – in which case there’d be no paradox to start with.

18 Thanks to Murat Aydede, Sara Bernal, David Christensen, Maite Ezcurdia, Hilary Kornblith, Shaun Nichols, Derk Pereboom, and Jesse Prinz for helpful feedback. This paper was first drafted when I was a postdoctoral research fellow at the Instituto de Investigaciones Filosóficas (UNAM) in Mexico City, the support of which is gratefully acknowledged.

REFERENCES


© 2002 University of Southern California and Blackwell Publishers Ltd.
THE PARADOX OF SELF-CONSCIOUSNESS REVISITED