The roots of self-awareness

Michael L. Anderson[1] and Donald R. Perlis[1,2]
[1]Institute for Advanced Computer Studies
[2]Department of Computer Science
University of Maryland, College Park
{anderson, perlis}@cs.umd.edu

Abstract

In this paper we provide an account of the structural underpinnings of self-awareness. We offer both an abstract, logical account—by way of suggestions for how to build a genuinely self-referring artificial agent—and a biological account, via a discussion of the role of somatoception in supporting and structuring self-awareness more generally. Central to the account is a discussion of the necessary motivational properties of self-representing mental tokens, in light of which we offer a novel definition of self-representation. We also discuss the role of such tokens in organizing self-specifying information, which leads to a naturalized restatement of the guarantee that introspective awareness is immune to error due to mis-identification of the subject.

1. The essential prehension

John Perry once noticed a trail of sugar on the supermarket floor (Perry, 1977). Thinking to himself, "someone is making a mess," he set out to find the person responsible and stop him. Thus, pushing his shopping cart before him, he began to follow the trail in search of the mess-making shopper while sugar continued to leak from the torn bag in his own cart. This amusing and self-defeating incoherence ends only when he realizes:

(1) *I* am the one making the mess.

Perry calls the "I" in this realization the "essential indexical"—essential because no belief of the form "X is the one making a mess" will cause a shopper to check his own cart *except* one in which X has, or can be made to have, the indexical and self-reflexive character of "I". Thus, for instance, the belief,

(2) The only bearded philosopher in the market is the one making a mess.

will only cause Perry to check his cart if he also believes that he himself (he* (Castañeda, 1966)) is the only bearded philosopher in the market, a belief he would express as

(3) *I* am the only bearded philosopher in the market.

For Perry, the main upshot of the discovery of this essential indexical is that one must make a "sharp distinction between objects of belief and belief states" (p.144); after all, in the circumstances discussed, both (1) and (2) have the same object, and the same truth conditions, but (2) alone (that is, (2) in the absence of (3)) is not sufficient to cause Perry to stop his cart—indeed, we can imagine him redoubling his efforts to find the right shopper, now that he knows he is looking for a bearded philosopher—whereas (1) will stop him right off. Although they may have the same object, (1) and (2) are clearly not the same belief. Thus, Perry presents the essential indexical as a special case of the more general, and well-known fact that it is possible in thought (and thought's various vehicles, e.g. language, belief, etc.) to express the same relation(s) between the same objects in ways different enough to coherently and unknowingly take divergent attitudes toward, or different actions in the face of, each expression. Thus can one affirm the cleverness of Cicero and the stupidity of Tully, hope one is elected consul and campaign against the candidacy of the other. Likewise, one can notice a trail of sugar on the floor of the supermarket, think that someone is making a mess, hope that (s)he will stop, but continue to push one's shopping cart, containing a leaking bag of sugar, all over the market. A great deal has been written about this phenomenon since Frege first confronted it (Frege, 1960) and we will not add further to the flood of ink here.

Instead, we would like to shift focus to an issue which has gotten somewhat less attention. This issue comes to the fore when one considers how, exactly, Perry came to know that it was he himself making the mess, and, more generally, how any person, starting from some given third-personal characterization or bit of information, could come to realize that the fact pertained to him or her self. As Perry makes clear, no description, however detailed and specific, will be sufficient to point the thinker to him or her self *as such* unless that thinker already has a grasp on the self of the right sort, with the right content. Thus, to generalize the point made already above, no belief of the form "The one making the mess is F" (where F is a definite description taking one's self as its object) will lead to the belief "I am the one making the mess" without the intervention of a belief of the form "I am F". This problem goes very deep indeed, for it implies further that one cannot come to believe anything of the form "I am F" (or, "The F is I") without, in each instance, a grasp on the self that is prior to one's awareness of any particular F. We seem put in a situation reminiscent of the paradox of the learner, in which we are forced to admit that before we can learn anything about the self, we must know it already. A solution to the puzzle appears to require a prior grasp on the self that is, to borrow an oft-used phrase, always already there, underlying all our substantial self-representations. Or, to put it in somewhat more familiar terms, there must be a grasp on the self which is *identification independent* (Shoemaker, 1968) in order to ground self-knowledge that requires self-identification. We call this grasp the *essential prehension*.[1] For reasons that will become clearer as the essay proceeds, we will argue that the

[1] The problem of essential prehension is to be contrasted with the paradox of self-consciousness recently identified by José Luis Bermúdez. According to Bermúdez' multi-lemma paradox, it is impossible to give a non-circular account of self-consciousness for (to express it briefly), one cannot analyze self-consciousness in terms of self-reference, that is, in terms of mastery of the linguistic pronoun "I", for proper use of this self-reflexive pronoun assumes self-consciousness. Yet, one cannot characterize self-consciousness except in terms of the ability to think a certain class of thoughts—namely, those canonically expressed with the use of the linguistic pronoun "I". Thus, self-consciousness apparently presupposes mastery of the first person pronoun (Bermúdez, 1998).

Unlike our puzzle of essential prehension, Bermúdez' paradox turns on the assumption that all thoughts (and thus by consequence "I"-thoughts) are necessarily expressible in language. Or, to put the point in the methodological garb which Bermúdez adopts: "To understand what it is to be capable of thinking a particular range of thoughts, one must first find the canonical linguistic expression for the thoughts in question and then explain the linguistic skills that must be mastered for the use of that linguistic expression." (p. 13) Thus, for Bermúdez, the path to a solution lies in identifying forms of awareness which deliver non-conceptual (and thus non-linguistic) first-person (self-representing) contents, allowing for the possibility of (at least some kinds of) self-consciousness without language.

In contrast, the problem of essential prehension cannot be solved by appeal to any kind of non-conceptual self-representation, for the grasp in question is precisely what allows *any* representation to be *self*-representation, what allows the identification of one's self as an object picked out by some arbitrary representational content F. And neither does this matter reduce to the general question of how objects are identified and individuated, that is, of how contents are attached (or experienced as belonging) to objects. For it is evidently possible to be in the position of having identified an object (picked out by a descriptive content F, e.g. "the bearded fellow over there" as seen in a mirror), and even to be able to provide further descriptive detail ("wearing the checked shirt"), and yet still fail to realize that the object (bearded fellow) in question is one's self. In so far as this is true, this failure cannot be explained simply in terms of (some failure of) the general ability to match objects with appropriate descriptive contents; this ability is fully functional in the example. Rather, the question is how *any* representational content gets essential prehension must be understood not in terms of a special kind of self-identification or privileged self-awareness, but rather as a process of self-establishment or stipulation, along with the structural facts that make such self-stipulation coherent.

It should also be noted that the problem of essential prehension in no way depends on whether the states involved are conscious; it is not a problem of *consciousness*, but a question about the basis for casting any representational or informational state in first-personal form. This is to say that the purpose of the current essay is to identify and outline a solution to one of the many “easy” problems of consciousness. Although we expect that the structures and mechanisms we identify as central to our solution do indeed help determine the shape of self-awareness in conscious beings, the current essay makes little direct contribution to the “hard” problem *per se* (Chalmers, 1995b).

2. Motivating knowledge without indexical knowledge

To emphasize the fact that the problem under discussion is not a problem of consciousness *per se*, and to make the analysis somewhat cleaner by sidestepping some of the more difficult issues of awareness, we will approach the matter first by re-casting the problem purely in terms of the kinds of informational or representational states which could be manipulated by a zombie (Chalmers, 1995a), or a robot. Thus imagine a robot, JP-B4, that is leaking oil. Such a robot might well see the oil on the floor, causing a “belief”[2] with the following content to be entered into its knowledge base (KB): “Some robot is leaking oil”. The questions for JP-B4 are, what additional information, beliefs, control structures, and the like, would it need to:

attached to (or any perceived object is identified with) the singular designation: *one’s self*. Of course, this doesn’t mean that Bermúdez’ discussions of the various forms of self-awareness are entirely unhelpful, as indeed we will be covering some of the same ground later in the essay, albeit with somewhat different ends in view.

[2] We employ scare-quotes here to acknowledge the fact that belief is a complex notion; see (Perlis, 1990).

(A) Come to hold the belief that it, itself was the robot leaking oil, and

(B) Bring itself (rather than, say, KQ-C5) to the repair shop.

Let us start with requirement (B) first, and thus assume that the following belief has somehow appeared in the KB:

(4) JP-B4 is leaking oil.

Under what conditions would this belief cause JP-B4 to go to the repair shop? Following the form of the Perry example, we appear to need an additional belief of the form

(5) I am JP-B4.

from which it would be possible to conclude:

(6) I am leaking oil.

But what, exactly, is the power of belief (6)? Perry's account of the puzzle of the leaking substance rightly focuses on the motivational difference, but suggests that this difference can be accounted for in terms of the presence in (6) of the indexical token "I". In our view, this is somewhat misleading, for it does not seem that "I" is either necessary or sufficient to account for this motivational difference.[3]

Imagine that JP-B4 has the following rules in its KB:

leaking_oil(x) → needs_repair(x)

(needs_repair(x) & ¬at_location(x, RepairShop)) → get_to_repair(x)

where *get_to_repair()* is a function which implements the following algorithm:

```
function get_to_repair(RobotName)
{
```

[3] It should be noted that, despite appearances, Perry does not think that indexicals—"I" in particular—are necessary for thinking self-thoughts. See, e.g., "Belief and Acceptance" in (Perry, 1993). Of this common misunderstanding of his argument in "The Essential Indexical" he writes: "…[T]he problem referred to in the title had to do with the fact that indexicals *seemed* essential to *expressing* certain thoughts; from this some readers seem to have assumed that I thought that indexicals *were* necessary for *having* those thoughts." (Perry, 1995, fn. 4). None of the arguments of the current essay depend upon, or encourage, the supposition that "I" is necessary for self-reference, nor that Perry thinks so.

```
  find(RobotName)
  grab(RobotName)
  tow(RobotName, RepairShop)
}
```

In such a case, coming to belief (6) will cause JP-B4 to try to find, grab and tow *I* to the repair shop. Most likely, it will report that it is "unable to find I". Indeed, it would be straightforward to implement a robot with exactly this (incorrect) behavior. What has gone wrong here? Must we say that for JP-B4 belief (6) is not appropriately indexical, does not, in fact, refer to JP-B4? If so, it is *because* it fails to play the right motivational role. Indeed, we can push the thought experiment to its extreme. Imagine that the function *get_to_repair()* looked instead like:

```
function get_to_repair(RobotName)
{
  if RobotName = "JP-B4" then
      go(RepairShop)
  else
      find(RobotName)
      grab(RobotName)
      tow(RobotName, RepairShop)
  endif
}
```

where *go*(*Destination*) is a function causing JP-B4 to travel to the assigned location. At the same time, let us imagine further that JP-B4 has a fully implemented language processing engine, which respects all the linguistic rules for "I". Thus, when reporting on its own states, JP-B4 uses the token "I", such that beliefs predicated using "JP-B4" are expressed using "I" (e.g. *leaking_oil(JP-B4)* becomes "I am leaking oil"), while other beliefs are expressed as predicated (e.g. *tired*(*Joan*) becomes "Joan is tired"). Likewise, JP-B4 properly interprets the use of "I" in others, so that "I am tired" is entered into the KB as *tired*(*CurrentSpeaker*).[4]

[4] We can even imagine JP-B4 is able to interpret its *own* utterances the same way, so that when it hears itself say "I am leaking oil" it concludes: *leaking_oil(JP-B4)*. Thus, were JP-B4 to have a belief in its KB of the form

For such a robot, it would appear that "I" has its normal linguistic properties, but none of the motivational ones, which are instead invested in "JP-B4". This suggests a minor puzzle for JP-B4: when asked why it is going to the repair shop, what ought it say? The linguistically correct answer is to say "I am leaking oil"; but if this were a direct report of a belief of the form *leaking_oil(I)* it would be false, and indeed were JP-B4 to have only such a belief, it would in fact have no reason to go to the repair shop. The answer "I am leaking oil" is correct only in so far as it is an expression of the belief *leaking_oil(JP-B4)*, and thus the motivation to which JP-B4 alludes with its use of "I" comes not in virtue of the linguistic function of "I", but rather in virtue of its internal connections to the representational token "JP-B4".

Whatever the right way for JP-B4 to justify going to the repair shop, and however the puzzle thereby raised is to be solved, the thought experiment as a whole strongly suggests, following Ruth Millikan (1990), that in addressing the puzzle of the leaking substance, one should focus squarely on the motivational issues, and let language fall where it may. The relevant difference between (1) and (2) and between (4) and (6) is that the former, but not the latter, predicates the salient fact of the symbolic expression which is, in point of fact, connected in the right way to the action-producing components of the agent in question.[5] For John Perry, and indeed for anglophones in general, that special symbolic expression is "I"; for JP-B4 it appears instead to be "JP-B4".[6]

leaking_oil(I) it could say "I am leaking oil" thereby entering this belief into the KB predicated using "JP-B4" rather than "I".

[5] There is an ambiguity here which it is worthwhile to point out, although we will not otherwise pursue it. The claim is that (1) and (2) differ in the functional role that they have for John Perry, and the focus has been on the impact it has on John Perry's *actions*. However, the two sentences differ also in the inferences they will permit. Thus the sentences differ not just in allowing or motivating different physical actions, but also different *mental* ones.

[6] NB: the lexical form of these expressions is, of course, arbitrary. We avoid the use of "I" for JP-B4 to emphasize that what is at issue is neither lexical form nor linguistic rules of expression, but rather the architecture of the system, and the role played in it by particular representations, whatever their form. In the architecture we describe, "I" could

3. Self-knowledge in JP-B4, a first look

It takes only a small amount of further inquiry to realize that requirement (B) suggests that the token “JP-B4” be implicated in a great deal more than just a special case of the function *get_to_repair()*. If, for instance, the robot is actually to *get* to the repair shop, it must know not just where the repair shop is, but also where *it* is. There are at least two general methods a robot could use to determine its location. It might track this information directly, by being equipped with a GPS receiver for instance, or it could try to use more general-purpose sensors (vision, laser range finders, etc.) to try to determine its location. In the former case, the most sensible design choice is to have the information delivered by the location sensor rendered directly in the proper form, e.g.: *at_location(JP-B4, x, y)*. In the case where more general sensors are used—sensors that deliver information about JP-B4 *other* than its location, and/or information about objects *other* than JP-B4 (including their location)—this expedient is of course not possible. Still, for the information to be useful in guiding JP-B4 to its destination, it must nevertheless be rendered in the same form. This suggests that JP-B4 requires a mechanism for *selectively* rendering some location information in the proper form. But here we would face another version of the puzzle of essential prehension with which we began, for on what basis would the selections be made? If the system provided information about the locations of objects in general, it seems that JP-B4 would have to already know which object it was, and from this determine its location. Thus, JP-B4 needs either a specialized system, to tell it only where *it* is, or a specialization of a general system, to somehow mark one of the located objects as itself. We will discuss this matter further when we turn to addressing requirement (A).

Still, we must keep in mind that any specialized system will not be able to give us other

of course be substituted for “JP-B4” in belief representation and language production, for whatever convenience this might provide.

information about the objects in our environment, and having such information is certainly necessary for a robot that could display the skills called for by requirement (B). For it cannot be the case that all the information gathered by the robot's sensors—*green(x)*, *leaking_oil(x)* and the like—can be organized under the token "JP-B4", lest it go alone to the repair shop even when it sees an oil leak in KQ-C5. Rather, it appears to be a general principle that any system that, for whatever reason, needs to respond differently to the same information depending on whether it pertains to itself or to something else—or even, for that matter, depending on whether it pertains to object A or object B, as for instance one might want to respond differently to the beliefs *coming_closer(Predator)* and *coming_closer(Mate)*—had better have a way of registering the relevant difference. For JP-B4 this means marking the difference with different representational tokens for each object in question, including, naturally enough, itself.[7] Thus, to meet requirement (B), JP-B4's informational system must be such that it can differentiate between different objects, appropriately sorting the information it gathers so as to render it in proper form: *leaking_oil(KQ-C5)*, *grimy*(*LR*-*D*6), etc.[8]

This general, perhaps obvious point can be interestingly extended to the case where a robot needs to make and maintain distinctions between not just the physical states or properties of multiple objects, but also the informational or belief states of other agents. Such an ability would be useful in cases where the activity of multiple robots needs to be coordinated, whether to allow them to engage in a cooperative task, or just to help minimize conflicts as they pursue their individual agendas. Whatever the precise situation, whenever a robotic agent maintains information about the belief states of other agents, it needs some mechanism for keeping the

[7] Of course, while there are many ways of allowing for the appropriate registration of the relevant differences here, it is perhaps worth noting that a representational system with a subject-predicate structure offers an rather elegant and flexible way of capturing the required information.

states appropriately sorted by their actual bearers. For an extreme example, imagine a team of search-and-rescue robots, each of whom has direct access to the perceptions and KB of every other member of the team. There are many ways in which such access would be useful, for instance, in allowing the team to pinpoint the location of a rescue beacon through triangulation, or to be able instantly to make inferences about evidence gathered at disparate locations. However, without some way of distinguishing the beliefs according to their primary bearers, contradictions and incoherencies would soon result. For instance, were the belief—initially held by JP-B4—that a train is rushing closer indiscriminately combined with belief initially held by KQ-C5 that it must be still to avoid a prowling tiger, the result might be a crushed JP-B4 (who stayed still to avoid the tiger), a mauled KQ-C5 (who ran to avoid the train), or at the very least two robots who believe that they need to run and stay still at the same time. A similar story might easily be imagined in the case where robots conflate the *perceptions* of other robots with their own, causing them to be guided locally by non-local perceptions. Such conflations and contradictions would quickly make coherent action impossible.[9] And, of course, the same problems arise even if the robots come to have such beliefs not through direct access, but through some kind of inference from observation. The natural solution is an extension of the one previously adopted: JP-B4's beliefs ought to have forms like *believes*(*KQ-C*5, *W*) and *believes*(*JP-B*4, *X*), and perhaps even *sees*(*KQ-C*5, *Y*) and *sees*(*JP-B*4, *Z*). Such tagging would ensure that all known information about each tracked object was properly integrated, and would trigger the appropriate responses.

[8] Once again, this ability includes the ability to appropriately assign information to "JP-B4", and once again the puzzle of essential prehension looms: on what basis is information assigned to the motivating representational token "JP-B4"? We will address this question directly when we turn our attention to requirement (A).

[9] This is not to say that one robot would never come to believe something based on the belief of another; but one would not want simply to adopt every such belief without consideration, and without rules for appropriately changing their form to reflect the fact that the belief is now to be held by a robot with a different KB, at a different location, and in a different situation.

4. Self-representation defined

We have argued so far that whatever information JP-B4 needs to gather, insofar as it needs to respond differently to that information depending on whether it pertains to itself, or to another robot, it will need individual representational tokens under which to organize that information. Further, it will need a special representational token for itself, which is distinguished from the others only by its *particular* motivational and representational role. This is a point worth emphasizing: the token "JP-B4" is not special in virtue of its general functions to organize perceptual information and beliefs and to guide action. "KQ-C5" has these very same functions. Both tokens organize information and guide action because each is a representation; *being* a representation consists in having such roles. Nor can these two tokens be distinguished by the detail or richness of the information that they organize. As we have seen, even the most immediate information about KQ-C5's perceptions and beliefs can be represented by JP-B4, and this information need not be of lesser quality than the information JP-B4 has about itself. Indeed, we can imagine an impairment of JP-B4's self-representations such that the information organized under "KQ-C5" is in fact richer and more detailed than that organized under "JP-B4". This fact would not in any way make "KQ-C5" a self-representation for JP-B4. Rather, the distinction comes down to a simple principle, which can be stated succinctly by using disquotation.[10] The token "JP-B4" is a self-representation for JP-B4 just in case:

(7) JP-B4 represents with the token "JP-B4".
(8) "JP-B4" is a representation of JP-B4.[11]

[10] In his paper "Self-notions" (Perry, 1990), Perry presents ideas that patrially anticipate some of our treatment here; the three criteria (7-8-9) that we offer for self-representation may come close to providing necessary and sufficient conditions for what Perry calls self-beliefs.

[11] This is to say, whatever the criteria for the act of representing, and for being a representation turn out to be, it is assumed for the sake of the argument that the token "JP-B4" represents JP-B4. Those criteria may well include not just gathering or containing information derived from JP-B4, or having an informational link to JP-B4 (and therefore, under the right conditions, co-varying with JP-B4), but also that the token in question is standardly used to guide

(9) Any transitive action, taken by JP-B4 and containing "JP-B4" as its direct object in the description under which JP-B4 takes the action in question,[12] will be directed at JP-B4 in actuality.

Thus, for instance, the description "The only bearded philosopher in the market", entertained by John Perry, meets criteria (7) and (8), for it indeed represents, and is being represented by, John Perry. However, it fails on criterion (9), since the intention to "Stop the only bearded philosopher in the market, as he is making a mess" would not cause John Perry to stop.[13] Likewise, for JP-B4 the token "I", even under those conditions where it fulfills criteria (7) and (8), does not, for JP-B4, meet condition (9). To get JP-B4 to go to the repair shop, the object of its described intention must be "JP-B4", for only this token is so integrated into the perceptual and control systems of JP-B4 to cause it to take JP-B4 as its object in action. That is the specific point; it is worth making explicit a more general one: in order for a token to have the right representational properties—e.g. of representing the self *as such*, under the proper "mode of presentation"—it must have the right motivational properties, or, to reverse the claim, it is partly in virtue of having certain action-guiding properties, a certain role for the representing agent, that a token has its representational properties (Rosenberg and Anderson 2004; forthcoming).

actions taken toward the object in question (Anderson, 2003b; O'Donovan-Anderson, 1997; Rosenberg and Anderson, 2004; forthcoming), as is suggested by criterion (9).

[12] That actions can be taken under different descriptions is an extremely old observation, and its consequences have been extensively examined in both philosophy and literature. Oedipus, to take a famous instance, killed an old man in the road who insulted him, and slept with the Queen of Thebes; his transitive actions were directed toward their objects under *these* descriptions, rather than the shorter, and more forbidding descriptions "my father" and "my mother". (For a short but thorough discussion of these matters, see (Anscombe, 1963).) Now, for JP-B4, given its design, taking an action under a certain description means literally that the intention that effects the action in question will contain the very token "JP-B4" in the direct object position. Considering the condition more generally, it requires only that the mental token or description named in (7) and (8) comprises the direct object of the description under which the action named in (9) is taken, however such a description is generated or intention is instantiated.

[13] Likewise, the intention to "Shave the only bearded philosopher in the market" would not result in John Perry being shaved; indeed, even were he to see the person under this description (perhaps in a mirror), the logical result of such an intention, guided by the representation but not taking its actual object as the object of the action, would be for John Perry to apply shaving cream to (and attempt to shave) the image in the mirror. Exactly such behavior has been observed in patients with certain kinds of brain damage: told to go into the bathroom to put cream on their face, they instead apply cream to the image in the mirror. (From personal communication with the Center for Neuro-Rehabilitation in Annapolis, MD.)

5. Self-knowledge in JP-B4, revisited

This returns us to requirement (A): granted that JP-B4 needs such structures as described to be able to bring itself appropriately to the repair shop, how is it that information comes to be organized under the proper representational tokens to begin with? How could JP-B4 have come to the belief "JP-B4 is leaking oil"?

Well, might not JP-B4 simply *see* that it is the one leaking oil? Can it not just look down, and, noticing the trickle of oil down its torso, conclude, "JP-B4 is leaking oil"? Perhaps. Let us grant perceptual abilities sophisticated enough to identify a robot torso and a trickle of oil, and inferential abilities sufficient to conclude from this information that a robot is leaking oil. But what warrants the identification of this robot as "JP-B4"? We briefly argued earlier that this matter could not be settled by appeal to the general ability to organize perceptual information into objects (and under object-representations). For after all, although there may well be perceptual clues that allow the identification of JP-B4 as an object, what in perception could allow for the determination that it is "JP-B4"? When we re-identify an object, we often do so on perceptual grounds (we *recognize* it), and we can, thereby, use this already formed representation to organize current perception. Naturally, if "JP-B4" already exists as a representation we can employ this same method—but of course, that is just a specific example of *already knowing* which thing JP-B4 is. Nor, of course, will the creation of representations *ex nihilo* (which might work with newly encountered objects) help in representing JP-B4, for information about JP-B4, to be properly effective, must be organized under that very representation of JP-B4 which meets criteria (7)-(9), above. Perception alone may be able to create representations which satisfy (7) and (8), but in order to satisfy (9) JP-B4 must be able to organize self-specifying perceptual information under that very representation which is implicated in JP-B4's action control systems so as to allow self-oriented actions.

In our earlier discussion of how JP-B4 might determine its *location*, we suggested that the problem called for specialized perceptual processing mechanisms, designed precisely to deliver information in the required form. The same might be true here, but let us take the opportunity to examine the issue more closely, for the most elegant solution to a specialized problem is one which follows naturally from the solution to one more general. Is such a solution available here? For the sake of the argument, let us assume that the full range of techniques which can be brought to bear on the general perceptual problem of parsing sensory information into objects—for instance, the principles of cohesion, contact, continuity, solidity, etc. (Spelke, 1990; Bermúdez, 1998)—can also be brought to bear on the problem of identifying JP-B4's body, which is, after all, also an object that JP-B4 could perceive. Thus, we take it as a given that JP-B4 can form a representation of its own body as one object among many, but what allows it to mark this particular body as its own? One possible answer has to do with the location of these various objects in perceptual space. An important function of perception, at least for any agent that could be called upon to physically interact with the objects it senses, is to determine the positions of perceived objects relative to the agent. As JP-B4's engineers, we know that the simplest way to solve this problem is by defining the coordinate system on which JP-B4 locates perceived objects in such a way that JP-B4 is at the origin. That way, every location given by perception is *already* a relative location. But the consequence of this is that there will always be one object identified by perception that contains that origin. A simple rule which requires information about the object containing the origin to be organized under the token "JP-B4", then, might serve as the basis for JP-B4's ability to appropriately organize self-specifying perceptual information under its self-representing token, and without requiring much in the way of specialized processing mechanisms. An oil leak seen by JP-B4 to be in the object containing the perceptual center, then, is only a step away from the conclusion “JP-B4 is leaking oil”.

Still, as elegant as this solution is it is also somewhat limited. The first limit has to do with the range of its application. It is possible to define the perceptual origin in the way suggested above only if the relation of the sensors to JP-B4’s body is known. If this relation is fixed (as it generally is, for instance, with a laser range finder), then of course, the relevant information can be built into the solution itself. But if this relation is variable, as it would be, for instance, for a movable camera or touch sensor, then placing JP-B4 always at the perceptual origin depends on determining the actually current relation between the sensors and JP-B4. Thus, solving the problem for a robot with sophisticated sensors requires another way of gathering this information—itself self-specifying, and therefore affected by the same puzzle as the rest. Second, any primarily exteroceptive system is going to be limited in its ability to gather self-specifying information in the first place. It is unlikely to be the case, for instance, that all of JP-B4’s body will be in JP-B4’s own sensor range. Further, if JP-B4 has movable parts the state of which it is important to track, it is not likely that applying an exteroceptive system alone—say vision—will provide the most efficient solution. Consider, for instance, JP-B4’s arm. In order to use this limb in even the most rudimentary way, it will be necessary to know, and be able to track, its current position. Surely it would not be wise for vision to be the sole means of doing this. For it might not be the case that the arm is always in view, especially if the object to be grasped is moving and therefore requires the visual system to move with it, even if that means losing sight of the arm. Indeed, tracking objects is quite a hard enough problem for a visual system without adding the burden of simultaneously tracking the position of a grasper. Further, it would be somewhat odd to have a robot which needed literally to *look* for its own arm, and, having caught sight of some arm, determine whether it was *its own* by seeing whether it was a part of the object containing the perceptual origin. The rule proposed above, while theoretically sound, appears to require significant supplementation.

In point of fact the two issues raised above can be dealt with in the same way, by allowing JP-B4 to have *direct* access to the locations of its movable sensors and effectors. It should not be surprising that it is standard practice to equip robots with just such an information feed; by the use of such devices as joint angle sensors, a robot is always in a position to know where its movable parts are. Note that unlike the case with exteroception, the general function of which to gather and track information about *many* objects naturally raises the question of how the identity of the perceiver itself can be known, this question does not arise for components like joint-angle sensors whose assigned task is to gather information about various aspects of the robot itself (and are known therefore as proprioceptors). As with the specialized system for tracking JP-B4's (objective) location, the information coming from such proprioceptors would naturally be cast in self-specifying terms—for JP-B4 that means, of course, that this information is represented in terms of (predicated using) "JP-B4". By design these sensors gather information about one and only one object, and the identity of that object can therefore be assumed when structuring and interpreting the information; there is no sorting problem in need of solution here.

Thus, we must imagine that JP-B4 has proprioceptive sensors, the outputs of which are self-specifying representations of the spatial position of its limbs and movable sensors. In the case of JP-B4's effectors, used to manipulate objects in the environment, that spatial position needs to be known relative to those objects. The issue can easily be addressed by representing the limb in the very same coordinate system used to represent the locations of objects. Alternately, the body may be tracked in its own special coordinate system,[14] which could be integrated as needed with the information tracked by other spatial-perceptual systems. Whatever the case, note the important implication that one of the things JP-B4 will know, as a result of its proprioceptive sensors, is where in the perceptual space its own limbs are.[15] On seeing its limb, then, there is no reason for the question of its identity to be raised: JP-B4 already knows what (and whose) it is.[16] Thus, were JP-B4 to see oil leaking from its elbow, it would need no further information or inference to conclude "JP-B4 is leaking oil".

This analysis could easily be extended to encompass robots with proprioceptive sensors of much greater sophistication. At the extreme, imagine that JP-B4 were covered with artificial skin, with multiple tactile receptors and damage detectors. For the tactile receptors to perform their exteroceptive task (providing information about objects in the environment) it will be necessary for the location of the receptors to be known, thereby allowing the stimulating object to likewise be placed in the robot's perceptual space. This would require knowledge not just of limb position, but also a spatial representation—in some coordinate system identical to or otherwise compatible with the system used by exteroceptive perception—of the entire sensory surface. Likewise for the damage receptors: although knowing just that sensor 5x9Wz had been triggered might be of use to a technician who knew where that sensor was, this information would be of little direct use to JP-B4 (should it, for instance, wish to apply a band-aid) unless it, too, knew where that sensor (and thus the indicated damage) was. What JP-B4, a robot covered in a receptive skin requires, then, is a spatial representation of its entire body. As already noted, information from JP-B4's proprioceptors is automatically expressed using the token "JP-B4". The thought here is that the spatial representation of JP-B4's body, proprioceptively updated, would be integrated with

[14] As Shaun Gallagher suggests is in fact the case with humans (Gallagher, 2003).

[15] Milner and Goodale (1995) present compelling evidence for the existence in primates of not one single, central representation of space, but rather many task-specific ones. It may be that a similar solution would be the best for JP-B4 as well. But the main point still holds: for each spatial representation implicated in the manipulation of objects, JP-B4 must know, on those coordinates, not only where the objects are, but where it and its limbs are.

[16] Vision still might have a role to play, of course. If a limb does not in fact appear to be where proprioception says it should be (or vice versa, a limb contiguous with the object containing the perceptual origin appears where it is not expected) this could indicate some kind of problem. Similarly, vision might be used in cooperation with proprioception to guide delicate limb movements, or each might be used to help calibrate the other.

spatially-organized perceptual information, so that information coming from the perceptual space occupied by JP-B4's body-representation would, likewise, automatically be organized under the token "JP-B4". By such mechanisms would an oil leak, felt by JP-B4 (through its receptive skin) or seen by JP-B4 in an object at a certain location, cause an expression of the following form to be entered into JP-B4's KB: *leaking_oil(JP-B4)*, thereby producing the desired behavior.

6. Information sorting and the essential prehension

The general puzzle with which this essay began turned on the fact that any given piece of information about the self might (for whatever reason) be organized under a representation that was not operating as a self-representation. It seemed in such cases that the only way to relate the information properly to the self was to find some kind of informational bridge—a fact already cast as a self-representation but also related to the representation under which the information in question was organized. Thus, knowing that "The only bearded philosopher in the market was making a mess", it takes the self-representing bridge information "I am the only bearded philosopher in the market" to relate the mess making to one's self. But this requirement raised the question: how could any information be organized under one's self-representation in the first place? We suggest that the puzzle does not turn on any linguistic competence, but can be solved by any information-gathering agent with an essential prehension of itself, a basic grasp sufficient to allow for the required organization of information under the appropriate self-representation.

In the course of discussing a robot-appropriate version of this puzzle, we determined that the essential prehension could be accounted for relatively simply by a representational convention for proprioceptive sensors which automatically casts information from these systems under the agent's self-representation, and some method of integrating this information with information coming from exteroceptive receptors. Understood this way, the essential prehension does not require any special self-identifying systems or modes. Instead, it consists just in rules guiding the operation of components of JP-B4 that would in any case be necessary to its operation in the world. Note, however, that the rules which constitute the essential prehension are not rules of *judgment*; it is not a question of being guided by principles, or using some set of criteria to determine whether some item of information pertains to the self or not (for in true judgment there is always the question of whether, or how, to apply the rule in any given case). Rather, the essential prehension is best considered a set of structural features of JP-B4's information gathering mechanisms that automatically *dictate* how information is to be sorted and organized. The information gathered by proprioceptive sensors, and the perceived features of that object containing the perceptual origin (or known to be one's self in virtue of one's proprioceptively required and updated spatial self-representation) simply *will* be organized under one's self-representation, whether or not this is sanctioned by one's better judgment.

For consider the case where you awake in a laboratory, with an evil-looking scientist standing triumphantly nearby. Glancing down, you realize to your horror that you now inhabit and control a robot body. The very coherence of the example depends on the notion that, despite your remembered body-image, and a set of beliefs about yourself (and perhaps even very strong beliefs about the impossibility of such brain/mind transplants) which in no way correspond to the currently apparent facts, you would still immediately identify the very unfamiliar robot body as the one you now inhabit and control. Indeed, the thought-experiment seems to imply that so long as one is connected in a certain way to a given body, then despite very strong beliefs to the contrary, it cannot but continue to seem that one is now associated with *this* body, with *these* boundaries, in *this* place and circumstance; despite the initial contradictions in one's belief set, from the time of the experiment forward, information gathered by and of one's new body in its new circumstances will constitute one's self-specifying information, and will (presumably) eventually replace all the old, now outdated beliefs and self-image with ones more appropriate to the new situation.

The case is similar for the opposite situation, where instead of one's body being replaced, one's self-image and biographical memory is tampered with. Here again, you might awake and think that something very odd must have happened, since you seem to recall being a 300-pound professional football player and bachelor, and now apparently you are a married, 120 pound woman with three children; but here again the coherence of the example—the very ability to think this thought—requires that you identify your (current) body/self not only *without* any identifying information, but even given a wealth of currently misleading and incorrect self-specifying information. Were this not true—were it not the case that the self is automatically identified as a condition for *receiving* self-specifying information, and not only as the result of *having* it—you would instead have to imagine that when you woke up feeling hungry, you would, in order to feed yourself, go looking for the 300-pound football player you think yourself to be. But this does not appear to fit the facts of self-awareness, nor does it seem likely that such a mechanism for self-identification would support robust real-world agency. And, in any case, it does not seem that a being with self-representations of this sort could ever solve Perry's puzzle of the leaking substance; it would always be in the position of looking for another to feed when it should be feeding itself.

We suggest instead that something like the mechanisms outlined above, by which certain information is sorted and organized under a special self-representation (self-representing just in virtue of its particular connections to the information-gathering and action-producing components of a given system) are what account for a system's ability to solve the puzzle of the leaking substance.

7. Somatoception and self-representation

Our discussion of the representational structures and capacities of JP-B4 has now led us full-circle, back to the human case. In the rest of this essay, we would like to discuss human self-representation in more detail, and we will argue that human agents employ mechanisms for self-perception and information sorting very much like JP-B4's. This leads us to suggest not only that these mechanisms can account for our ability to solve such problems of first-personal knowledge as the puzzle of the leaking substance, but that the somatoceptively-grounded self-representation thereby employed, rather than any linguistic indexical, might form the basis for the representation of first personal knowledge more generally. In so far as it is able to accomplish this latter goal, the essay will provide detailed support for the claim (and an explanation of the fact) that a representation of the self necessarily accompanies every mental content in introspective reflection, although the current essay is neutral on whether the fact of this accompaniment manifests itself in experience in terms of a self-quale.[17]

Somatoception, the awareness of one's own body, involves many specialized sensors arranged into several distinct information systems. Thus, for instance, the sense of touch involves specialized receptors for detecting pressure on, or deformations of, the skin,[18] a different set of sensors for thermal reception, and two further sets for pain reception. In addition, we have systems providing information about the body's interior (allowing us to feel our racing heart or poor digestion), the vestibular system which provides information about the orientation and motion of the body as a whole, and the proprioceptive system which provides information about

[17] This latter claim was made in (Perlis, 1997).

[18] There are four identified classes of mechanoreceptors, determined by the size of their receptive field and the speed with which they adapt to a sustained indentation. The small-receptive-field, slow-adapting mechanoreceptors (SAIs) have been implicated in the perception of texture (Craig and Rollman, 1999).

the position and motion of the limbs in particular.

Despite the existence of multiple specialized systems, it is clear that the types of somatoception must cooperate in various ways, and also with other categories of perception. Thus, knowledge of the position of one's limbs can be given by proprioception, but also by touch (the feeling of the desk pressing against one's knees) and by vision (Ghez et al., 1995). Indeed, vision can sufficiently confound one's sense of limb position that it is apparently possible to locate—to feel—the touch of a feather in a clearly visible and strategically placed rubber arm, instead of in one's actual arm, which is being simultaneously touched but is hidden from view (Botvinick and Cohen, 1998). Likewise, a single touch can simultaneously give interoceptive information (a heat in one's finger) and exteroceptive information (the heat of the stove one is touching).[19] And finally, it seems that certain kinds of tactile perception, e.g. feelings of texture, insofar as they involve not just contact between the sensing organ and the object, but also the motion of that organ, require both proprioceptive and tactile awareness.

Of course, it is not the task of the current essay to provide a comprehensive account of somatoception, but rather to inquire into the roots of self-awareness. Thus, we will be focusing our attention on just the self-specifying aspects of somatoception, and on touch and proprioception in particular. It is our claim that these senses have a particularly important role to play in grounding the sense of self—that is, that they do for human agents very much what the equivalent senses did for the robot JP-B4. Consider, first, the most basic point, nicely laid out by Bermúdez:

> One of the distinctive features of somatic proprioception is that it is subserved by information channels that do not yield information about anybody's bodily properties except my own.... It follows from the simple fact that I somatically proprioceive particular bodily properties...that those bodily...properties are my own. (Bermúdez, 1998, p. 147)

[19] Nicholas Humphrey (1992) makes much of this duality in his own explanation of consciousness.

The claim, no doubt familiar to many, is that certain information channels—those which by their structure deliver information solely about one's own properties—can be treated such that all the information flowing from them is automatically tagged as pertaining to the self (however such tagging is to be effected). But granting that this information is self-specifying raises the question: what about the self is thereby known? What, for instance, does an agent know in virtue of feeling an itch?[20] He knows, first of all, that he (he himself) itches, and to know this it is not necessary that he make any judgment to determine whether it is he or someone else who itches. The feeling of an itch is quite unlike, for instance, seeing a bit of red or hearing a shout in the distance, and wondering who, exactly, is wearing red or shouting. The itch comes already associated with the self in experience, and the suggestion is that this fact is to be explained by the very structure of the information channel involved; it does not appear to be possible to know of an itch by feeling it, and simultaneously to wonder who thereby itches.

But one generally knows more than this, for one knows *where* one itches—where in what body part, where in one's action space, and where in shared (public) space.[21] It is such information, after all, which is required to direct a scratch, whether one's own or that helpfully supplied to one's back by someone else. This location information appears to be likewise an ineliminable part of the experience (one cannot say that one itches nowhere, without casting doubt on whether one is itching at all), and likewise constrained in its content. For one can have an itch only in a location apparently occupied by one's body; something experienced as being at a

[20] What follows is a paraphrase of the much longer arguments to the same effect given by O'Shaugnessy (1980) ch. 6, and bears some resemblance to the arguments given by Bermúdez (1998) ch. 6, who largely recapitulates O'Shaugnessy.

[21] This location information isn't always very precise (an itch can seem to be *generally* in the arm, but somehow hard to locate exactly), nor is it always specific (one can seem to be itching *everywhere*). But it does seem that *wherever* one is itching, it must be *somewhere* in one's own body. Further, one needn't necessarily know where an itch is in *objective* space, for this relies on knowing where one's body is in objective space (say, Connecticut), and one's sense of objective location can be impaired (as might happen after being on an airplane for some time) without necessarily affecting one's judgment of the location of an itch.

point three feet above one’s left shoulder could not qualify as an itch.

On the other side of the coin, the various psychological disturbances of the body (e.g. phantom limb phenomena and the feather-touching-a-rubber-arm experiments (Botvinik and Cohen, 1998) show that the felt location of an itch needn’t be in one’s *actual* body. Does this sever the close connection between the body and experience argued for above, and with it the connection being drawn between certain classes of experience and self-specification? Not at all. In fact, these phenomena seem instead to underline the fact that certain feelings are by their nature cast in terms of a body experienced *as one’s own*. Consider first that the feather-touching-a-rubber-arm experiment does *not* show that an itch can be experienced anywhere, or in any kind of space at all; the experienced tickle still seemed to each subject to be in his or her own arm. What is severed in this experiment is *not* the link between experience and the self, or experience and the body, but that between the actual location of one’s arm (an arm experienced as one’s own) and its felt location, as (misleading) visual information about the limb’s location trumps proprioceptive information. Phantom limb phenomena seem likewise to confirm this experience/body-space connection, in the opposite case where (misleading) proprioceptive information about the location of a limb trumps visual information. As is well known, a phantom limb is sometimes felt in cases of amputation, perhaps as the result of continuing signals from nerves that previously carried stimulations from a part of the arm no longer present. The fact that stimulation from such a nerve is experienced in terms of a seeming limb, of definite spatial extent and even of specific posture, strongly suggests the existence of an information processing requirement, imposed by the nervous system, that these signals must be interpreted in terms of a specific part of a specific body. And the fact that the limb thereby experienced is not experienced as alien (despite very clear and compelling evidence that it cannot be one’s own limb that one is experiencing) further suggests the close connection (likewise imposed by the information processing systems involved here) between bodily experience and self-specification. Indeed, the connection here identified is so close that one will experience unfamiliar or even impossible versions of one's body-space if the self-specifying information processing system requires it. Thus, for instance, if a subject grabs her nose while the tendon in the wrist is vibrated, the vibrations will cause the wrist to feel as if it is bending away from the face, and, because the hand is touching the nose, the nose will seem to grow (Ehrsson, 2002). Odder still, patient E.P., who has a congenital lesion of the corpus callosum and a frontal lobe lesion caused by aneurysm repair surgery, occasionally experiences a third left arm and even a third left leg (Hari et al., 1998). In the case of E.P. the phenomenon appears to be caused by a kind of proprioceptive "memory"; new proprioceptive information does not erase or completely update old information, so that when the left arm moves it at the same time seems to be where it previously was. E.P.'s information processing systems are constrained to interpret proprioceptive information in terms of specific parts of a body experienced as her own, and they therefore meet this requirement by interpreting the doubled proprioceptive information in terms of a third arm or leg.

To account for these various phenomena we follow O'Shaugnessy in suggesting that touch utilizes a spatial *representation* of the body, not identical with the body itself (although doubtless largely determined by it), onto which such things as itches are projected, and in terms of which itches are therefore felt.[22] This seeming body—called the body-image by O'Shaugnessy, but better known as the body schema (Gallagher, 1986)—might be considered to consist of all the places at which an itch might be felt to reside. It appears as though the sense of touch *requires* such a body schema to account for the content of the experiences it delivers, in both the normal

[22] That the brain maintains precise somatotopic maps of the body is fairly well established, and the experienced location of tactile stimulus can of course be explained in terms of the processing of the stimulus by these somatotopic maps. Note however, as with the question of perceptual space, there needn't be a *single* somatotopic map, and indeed there is evidence for simultaneous spatial coding of tactile information by different areas of the brain, presumably each for its own special purpose (Nicolelis et al., 1998).

and aberrant cases. Indeed, the phenomena discussed above suggest not just that a certain class of perceptual stimulation—itches, tickles, pains, proprioceptive seemings—is required by one's perceptual processing mechanisms to be experienced in terms of (as being in and of) a *seeming* body (the body schema); in addition, the experience is required—despite knowledge and judgment to the contrary—to be *self-specifying*. The itch, and the body in which it in experienced, necessarily seem to be one's own. It appears that these two conditions are mutually necessary for this class of perceptual stimulation: one cannot experience an itch in a body which does not seem to be one's own, nor can one experience an itch as one's own which does not seem to be in a body. This mutual necessity suggests that, at this very basic level, self-representation is bodily-representation, and the self is known as, and in terms of, its body.

Thus, it appears that, as with the robot JP-B4, the senses of touch and proprioception (in cooperation with other forms of somatoception) equip an agent with a spatial sense of his own body, its general shape, and current disposition, cast (among perhaps many other forms) in spatial terms compatible with his action space and exteroceptive visual space. Such self-representations, required for self-maintenance (in which the somatoceptive senses are crucial) and in order to perform even such simple actions as reaching for an object (one needs to know both where the object is, and where one's hand is), we call *physical* self-representations. Although there is some debate about the prominence which should be given to representation (especially symbolic representation) in any explanation of human intelligence (Anderson 2003a; 2003b; Brooks, 1991; Chrisley, 2003; Kirsh, 1991), and although clearly not everything delivered by the various forms of perception is available for explicit symbolic representation (Edelman, 2002), for the purposes of the current essay we will be assuming that there is sufficient advantage to, and evidence for, structured representations[23] generally construed, to warrant the claim that there is a great deal that *is* so represented, including much of what is perceived about one's self. Thus, let us hereby introduce the representing token "SR*". Our proposal is that our information systems are set up such that information coming from somatoception is automatically tagged with such a mental token "SR*", and that "SR*" is a self-representation according to the criteria provided earlier:

The token "SR*" is a self-representation for agent *A* just in case:

(7a) *A* represents with the token "SR*".
(8a) "SR*" is a representation of *A*.
(9a) Any transitive action, taken by *A* and containing "SR*" as its direct object in the description under which *A* takes the action in question, will be directed at *A* in actuality.

Thus, consider that agent *A*, by the mechanisms outlined above, comes to know: "SR* itches", and, as a result, forms the intention to "Scratch SR*". Does "SR*" indeed qualify as a self-representation for *A*? It seems so, for it meets criterion (7a) by hypothesis, and there seems little reason to doubt that it also meets criterion (8a); insofar as "SR*" is organizing (contains) information obtained from *A*, through a causal link with *A*, that is generally true of *A*, that "SR*" thereby systematically co-varies with *A*, and that it is used to guide *A*'s behavior with respect to *A*, then "SR*" has at the very least a very strong claim to be a representation of *A*. Further, in light of the discussion above, it seems that it will also meet criterion (9a). For part of the information that *A* represents with "SR* itches" is the *location* of the itch. Assuming only that *A* is able through normal motor function to direct a scratch (among other actions) at a location, then the scratch in question will indeed be directed at *A*. That a representation would qualify as a self-representation in this way is far from automatic. For consider, instead, that the agent came to know "The only bearded philosopher in the airport itches", and that the agent does not know that

[23] Here and henceforeth, the terms "representation", "representing", and their cognates should be understood in light of the very general thesis that a representation is any cognitive state of an organism, standing in for something else, and useful in guiding behavior with respect to that thing, regardless of how that state is instantiated (Rosenberg and

he, himself, is the only bearded philosopher in the airport. Here the agent may well intend to "Scratch the only bearded philosopher in the airport"; but to do this he would have first to "Find the only bearded philosopher in the airport."[24] Even assuming for the sake of argument that the definite description qualifies on criteria (7a) and (8a), it will fail on criterion (9a), for in each case the actions would be directed *not* at the agent, but outward, manifested perhaps in the activity of wandering about looking for a bearded philosopher.[25]

To be clear, what is required of *A* is that he possess not just the representing token "SR*", and the information gathered under it, but also the general ability to be guided by such information not just in acting *with* his body (e.g. knowing the current position of one's limb is a necessary starting point to any effective reaching motion) but also *toward* his body. However, it seems that such abilities are necessary for an agent to coherently or effectively act in the world *at all*. Consider an agent's need to reach to a point directly to his side at shoulder height. If the agent (correctly) represented his arm as being by his side, but, unable to use this information, chose to act as if the arm was straight in front of him, the result, rather than the required raising of the arm sideways, would be the swinging of the arm backwards (assuming the arm moved at all—the mechanics here are not simple), and in any case the failure to meet the need at hand. Likewise for an agent unable to use his representations of objects to choose and guide his actions—representing a hat in front of him, but unable to use this information to select and guide an action, he might choose to reach for the hat sideways, or to eat the hat like a pancake. Insofar as this is so, it seems safe to assume for agent *A* the ability to be guided by his representations.

Anderson, 2004; forthcoming).

[24] One of the effects of building the self-representation on somatoceptive information is that the object thereby represented is always present to the representer, not needing to be found. The object represented as the self is always within the immediate action space of the agent, and the agent is aware of this fact just in virtue of the content of the information organized under the self-representing token.

Given this admission, it would be equally straightforward to handle other like cases of self-representation, e.g. "SR* is hungry" or "SR* is bleeding".[26] For the information organized under "SR*" tells *A* where, and in what position he is; in virtue of "SR*" *A* knows the spatial extent of his body, and the current position of his limbs (not to mention his apparent orientation to the outer environment). Thus so long as *A* can use the content of a predicate-representation to choose an appropriate action (in response to "itches" he chooses scratch, in response to "hungry" he chooses feed, and in response to "bleed" he chooses bandage) and can use the content of the object-representation (its location, extent, orientation, shape, etc.) to guide the chosen action to its intended object, then it seems that any information organized under (predicated using) the token "SR*" would qualify as self-knowledge, and further that predicating this information using "SR*" would be sufficient to motivate appropriate, self-oriented actions. "SR*", then, plays the same role for *A* that Perry claimed for the essential indexical; it is the representing token in terms of which an item of information must be expressed to motivate appropriate action; and it is sufficient, for a bit of information to be treated as self-knowledge, for it to be predicated using "SR*". Thinking "SR* is hungry" or "SR* itches" or "SR* is bleeding" and knowing, in virtue of "SR*" where he, or his itch, or his cut is, *A* can (and presumably will) feed, scratch, or bandage

[25] It is no use objecting that *A* can't *see* whether someone is a philosopher. Evidence gathered at the APA conference in Boston, held in two hotels attached by a shopping mall, and therefore requiring the attendees to frequently roam said mall among many hundreds of "normal" people, suggests that one *can* tell the philosophers by sight.

[26] Note that we are not commited to the claim that there is one *single* token "SR*" which centralizes all somatoceptively gathered information under one representation. In a modular mind, information is only selectively integrated, and while it *may* be that one of the somatoceptive systems is *fundamental* in the sense that it produces an *original* "SR*", which is thereafter shared with other modules and processes, it needn't be this way for the system as a whole to work as described. The idea is rather that somatoceptively gathered information is automatically gathered under *at least one* such representing token with the properties (7a)-(9a); any such self-representing token is "SR*". Whenever such self-specifying information is integrated across modules or processes, it is likewise organized under an "SR*" token, which, insofar as it meets criteria (7a)-(9a), is for practical (functional, behavioral) purposes the "same" token. The issue of what constitutes "sameness" for a representing token in a modular mind is an interesting one; although we address it a little further, below, it certainly deserves a more thorough treatment than it will receive here.

himself, without the need to think any indexical thoughts.[27]

8. Spatial self-representation and the "I-context"

Having discharged the an important argumentative burden of the essay by showing that somatoception can indeed ground a self-representation sufficient to organize self-specifying information, and to guide and motivate action, thereby giving any agent so equipped the ability to solve puzzles of the general form of Perry's problem of the essential indexical *without* the need to think indexical thoughts, we hereby proceed to the further speculation that this self-representing mental token "SR*" can provide the basis for organizing other, and higher-order, self-representations.

Consider, for instance, the case of seeing one's hand and knowing it to be one's own. How might this be possible? One, but not the only (Milner and Goodale, 1995), important function of

[27] Before moving on, it is worthwhile to raise an objection to this account which may have occurred to the reader: given that *A*'s spatial representation of his body plays such a central role in guiding his actions, isn't it the case that an indexical has been tacitly assumed? In feeling at itch, or directing a scratch, isn't *A* thinking: "SR* itches (or wishes to scratch) *here*"? According to the objection, we haven't gotten rid of the need for indexicals; we have just replaced "I" with "here". It is true, of course, that a central feature of the guidance of appropriate action is the perceived *location* of the item (whether itch, cut, bowling-ball or hat) at which the action is directed. The objection is misguided, however, in apparently assuming that this location information is presented to the agent in indexical form. This needn't be the case, and, indeed, at the sub-conceptual level surely cannot be the case. For consider that any given indexical is a concept requiring of the thinker who deploys it that she master a certain set of conditions or rules which govern that deployment. In the case of indexicals this involves recognition that the same concept can be deployed in many different situations, and that the particular reference of the concept on any given occasion depends in a systematic way on the relation between the referential rules for the given indexical and the situation in which it is deployed. In the case of "here", the concept always refers to a particular (although more or less broadly defined) location, that, being a location, could also be identified in other terms. Thus, first of all, the mere fact of *A*'s fixing on a location doesn't require that the location be fixed indexically: "two inches above the elbow" will do just as well. Further (and in answer to those who are thinking that what is really represented in this case is "two inches above *my* elbow", thereby introducing the first-personal indexical token) the location need not be represented *conceptually* at all. Part of the content of *A*'s perception of an itch may be its location, cast not in terms of any spatial or geographic *concepts*, but just in terms of a specified location on *A*'s bodily representation, the content of which can be cashed out entirely in terms more appropriate to motor-control programs than to inferencing mechanisms. Indeed, it may be that, *were A* to choose to refer to a given location with the indexical "here", it is just such sub-personal, motor-oriented specifications of location which would give that concept its particular content on its particualr occasion of use. Thus, far from the ability to think about a location requiring indexical thoughts, it is rather *A*'s ability to represent a location sub-conceptually, in a motor-space suitable for directing action, that is required for thinking indexically. We suggest a similar story, below, for the indexical "I"; rather than self-identification resting on the ability to employ the concept "I", employment of the concept "I" rests on prior self-stipulation, of the sort we have outlined here.

vision is to deliver for cognitive processing a representation of "what is present in the world, and where it is" (Marr, 1982). It is generally supposed that this particular visual task involves a retinotopic map in the primary visual cortex, and proceeds through processing by stages, ending with a sense of what is where in the world, cast in terms of concepts and abstract symbols arranged suitably for such things as inference and planning. One stage of this processing, worth noting in particular, is hypothesized to involve deictic pointers (not yet abstract symbols) attached to the various thises and thats differentiated by object-extraction processes. At this stage the perceived objects are distinguished primarily by their location, and not yet by their class, category or other properties (Ballard et al., 1997; Carrozzo et al., 1999; Gallistel, 1990; Hurford, 2003; Marr, 1982; Milner and Goodale, 1995). Our suggestion is that the somatoceptively grounded body-scheme provides a spatial "self-context" which can be integrated with this spatially organized visual information; just as the self-specifying information coming from somatoception is automatically tagged with the self-referential mental-token "SR*", perceptions located within the spatial "self-context" are likewise tagged with this same token. Of course, it is something of a mystery how representing tokens of this sort actually work in the brain, and, assuming the plausible notion that the brain is largely modular and encapsulated (Fodor, 1983), there is some question about what it could mean for the "same" self-representing token to be used to organize information provided by different modular systems (vision, somatoception, practical reasoning). Still, consider that it is apparently possible to *recognize* someone, *recall* stored information about them, and, while maintaining a perceptual-informational link with them, note such things as what they are wearing, how heavy they've become, that they sound congested, and that they seem tired and distracted, all of which information can later be recalled with the rest. Our claim is only that whatever sense of "same" can be applied to the representing mental token in such cases, which has to organize stored and current information, as well as information coming from different sensory modalities and very different processing subsystems (e.g. very different information processing subsystems are likely involved in the detection of the properties "heavy" and "distracted") can also be applied to the case of the self-referential mental token used to organize somatoceptive information, self-specifying perceptual information, and even, as we will argue below, intentional and self-reflexive information. Indeed, if the somatoceptively grounded self-referring mental token "SR*" (and, more importantly, the self-stipulation on which it is grounded, and which it signifies) can be shared[28] with the visual system, we see no reason to suppose that it cannot be shared with any system, perceptual or cognitive, which is in the business of forming representations pertaining to individual objects, including the self. One particularly important instance of this, which we shall not discuss in any detail, would be the sharing of the self-referring mental token with the language module. That the language module would depend on, and respect, object identifications—including self-identification—provided by other mental mechanisms, rather than possessing and imposing its own special way of identifying these objects, is an assumption which accords with the general principles on which a modular mind is supposed to operate. Thus we will suppose (for no other reason than convenience of expression, for nothing crucial in the current essay depends upon its truth) that the self-referring mental token, grounded by somatoception and used by exteroceptive perception, is likewise used by the language module, where it is expressed using the linguistic token "I", "yo", "Ich", "je" or whatever is dictated by local convention.[29]

Thus, an agent equipped with a somatoceptively grounded "self-context"—which for

[28] We talk of sharing for simplicity, but we do not thereby suggest any literal passing of a representation from place to place within a brain, whatever such a thing would mean. How the integration of information from disparate mental systems under the same mental token might be effected is beyond the scope of this paper. What we *do* rely on is the assumption that some such integration can occur, and thus we assume it can occur with the somatoceptively grounded self-referring mental token as well.

convenience, and in accord with the above assumption regarding the operation of the language module, we shall hereafter call the "I-context"—can form beliefs about its own physical state as the result both of direct input from somatoception (e.g. I am hungry, I itch, My arm is bent, etc.) and also from exteroceptive perceptual input derived from that part of the perceptual space known to be occupied by the body (by either of the two mechanisms suggested above[30]), and therefore likewise organized under the same "I" (e.g. I am bleeding from a cut in my torso). Naturally, assuming a properly functioning perceptual system, it will also be able to form like beliefs about the physical state of the other objects it perceives.

We have argued so far that somatoception is responsible for grounding a mental token, under which self-specifying bodily information is organized, and which can also used to organize representations generated by other perceptual and cognitive systems. Note however that we have so far discussed only the self-representation of physical properties, known in virtue of an agent's perceptual connection with a particular object—his body. But naturally self-representation encompasses other kinds of properties besides the physical, among them intentional and self-reflexive. Intentional self-representation is, as the name implies, concerned with the ability to represent information about the intentional states of the self such as belief, desire and intention. Whereas at the level of physical self-representation the self is represented primarily as a body, at the intentional level the self is represented as an agent. Self-reflexive representation, on the other hand, involves representing the self as *representing*, and may allow for the represented *unity* of the self, the combination of all the self-specifying representations. Just as we have suggested above that visual information about the self is known to be self-specifying insofar as it, too, is

[29] Carruthers' suggestion that one of the functions of the language module is precisely to integrate the information generated by different modules and processes (Carruthers, 2002) is fully compatible with this account of the grounding of "I".

organized under the somatoceptively grounded mental-token, so we will argue that intentional and reflexive self-representation is the result of the use of this same token by the cognitive system charged with generating representations of mental entities like belief.

9. Higher-order self-representation

Let us first make the obvious point that an agent representing beliefs, desires or other intentional states that are in fact its own, is not the same as that same agent representing such states *as* or *under the description* “its own”. As with any other attributable property, beliefs, desires, and the like take objects in thought, and the question before us is how those objects come to be assigned, particularly in the case where the object in question is one’s self.

Of course, to some degree this question is bound up with the larger issue of how intentional states come to be represented *at all*. There are several competing accounts of this, in three major categories: the ‘theory-theory’ of mind (Baron-Cohen, 1994; 1995; Leslie 1994; 2000; Scholl and Leslie, 1999), the ‘simulation theory’ of mind (Carruthers, 1996b; Gordon, 1986; 1995; Goldman, 1989; Heal 1986; 1998a; 1998b) and the ‘primary interaction theory’ (Gallagher, 2001). Briefly put, the theory-theory of mind is just the idea that our understanding of mentalistic notions, and our ability (and tendency) to interpret agents in intentional terms, is the result of our possession of a theory of mind (ToM)—a set of folk-psychological concepts and the criteria for applying them to their appropriate objects. In contrast, the simulation theory argues that we instead use our own minds as an internal model of the other, and run off-line, counter factual simulations (e.g. were I to believe X, I would do Y; *A* is doing Y, therefore (possibly) *A* believes X). Finally, the primary interaction theory argues that “we have a direct, pragmatic

[30] Recall that, insofar as perception uses location-based deictic pointers for one sort of categorization of the objects it differentiates in the perceptual scene, perception has its own way of distinguishing the self from other objects, for it

understanding of another person's intentions because their intentions are explicitly expressed in their embodied actions" (Gallagher, 2001: p. 86). This current essay is no place to examine the relative merits of each account, nor to decide between them. Instead, allow us to take a step back and make a few general observations about what they have in common.

In each case, what is being identified and examined is a certain mental or cognitive ability: the ability to interpret agents intentionally. Thus, in so far as it is correct to say that, in general, mental abilities are made possible by mental modules, intentional interpretation should also be encapsulated in a module or modules for the same kinds of reasons: e.g. developmental arguments to the effect that the (implicit) theoretical knowledge or interpretational structures apparently required to explain the observed behavior could not be acquired by the age at which the behavior is observed (there are many versions of this poverty-of-stimulus argument, most famously deployed in arguments for an innate grammar (Chomsky, 1965; 1979; Pinker, 1994), and computational arguments which purport to show that a non-modular mind would be computationally intractable (Fodor, 1983).

In light of such considerations, let us posit a generic intentionalty module (GIM),[31] which operates according to a logic that may, or may not, be captured by one, another, or some combination of the suggestions made in the above theories. Although it is unclear, then, exactly *how* GIM would work, *what* it does would be the same no matter which of the competing theories one favors: it takes as input the (perceived) behavior of agents in given situations (where "behavior" is to include facial expressions, apparent direction of gaze or object of attention, and other appearances) and gives as output structured representations[32] of the intentions and/or

is the one object which always contains the spatial-perceptual origin.

[31] Although GIM recalls the 'intentionality detector' (ID) from (Baron-Cohen, 1995), its function is meant to be much broader and more generic.

[32] *See* fn. 22.

intentional states of agents, which can be used for guiding our own actions and responses. Whether or not these representations are rendered *conceptually* or *symbolically* will depend on the general conceptual abilities, and the current needs, of the representing agent in question. Although it is part of the function of the module as hypothesized to identify which objects in the world are agents (they are those for which GIM gives intentional representations[33]) it is no part of it to identify or label objects *per se*. Thus note that the object-identifications in question would have to be provided to (shared with) GIM, from whatever source such object markers come (ego-centric spatial coordinates, deictic pointers, linguistic names, etc.). The idea is that we are not generally aware of unattached intentional states, which are then sorted by object, but rather that among the attributes that can be detected/perceived in an already identified object are its intentional states.[34] GIM operates, then, in a manner analogous to one kind of perception: it applies categories—in this case categories from inter-subjective interpretation and/or folk psychology—to incoming information, producing structured representations useful in guiding action and (perhaps with some further processing) supporting inference.

Let us say immediately, to nip a potential objection in the bud, that it is no part of this hypothesis that GIM plays a role in *generating* intentional states, only that it plays a role in

[33] That it is both natural and compelling to attribute intentional states to certain objects (e.g. software agents, or patterns in cellular automata), and that it remains so even when we have strong theoretical beliefs about the inappropriateness of such attributions, might perhaps be taken as one sign of a sub-personal mental module at work—just as the Müller-Lyer illusion, in not yielding to any knowledge of the equality of the lines, suggests a cognitively impenetrable, sub-personal component in vision which is so structured as to cause in this case a perceptual seeming of unequal length.

[34] This is important for two reasons: first because it implies that GIM must respect the labeling and sorting done by other modules (which of course must be the case with any property-specifying modules; no module which would lose track of the object to which a property applies would be of much use), and second because it suggests something striking: that we routinely attribute "mental" states to *physical objects*. Note that this is not the claim that mental states are physical states—indeed, part of the purported need for GIM is to pick up where ToBy, a theoretical module which attributes physical properties to objects, leaves off (Leslie, 1994). GIM is not attributing mental states to objects *as* physical states, and indeed it need have no stake in or knowledge of the *cause* of, or underlying explanation for, the properties which it detects any more than ToBy's application of concepts like weight and force need be rooted in some understanding of atoms or electromagnetism; but its function requires that it treat (some) physical objects as the appropriate bearers of mental states.

representing such states in a form suitable for consumption by various action-guiding and inference-supporting mechanisms. Note further that this general hypothesis should not be construed as prejudicing the question of what the mind in general, nor mental states in particular, *are*, how they are instantiated, or how best to characterize them. A given cognitive state may be a distributed brain state, a language-like symbolic structure, a bodily disposition, an intentional state of the organism considered more broadly, or all, some, or none of these things. The GIM hypothesis is only that there exists a module or modules that, consistent with the principles of functional specialization, generates structured representations of the intentional states of agents, of the right form for use in guiding appropriate responses in light of the intentional and contextual background of the representing agent. Such responses can range from the near-reflex of pulling back from anger or aggressiveness, to the explicit consideration of motives, circumstances, and likely outcomes. GIM is hypothesized to be one of a set of systems aimed at producing implicit or explicit predictions of, and possible explanations for, the changing states of our environment and its objects. As with the systems that generate expectations regarding the movements of physical objects as such, predictions regarding the behavior of agents might in some cases be best characterized in terms of non-conceptual know-how (e.g. moving to the right position to catch a ball), and in others in terms of symbolically expressed know-that (explicit, linguistically expressible expectations for an object's path and likely effects). Thus, although the three classes of theory mentioned above differ in their commitments regarding what form the majority of intentional representations and their attendant expectations are likely to take—with the theory-theory favoring explicit conceptual representations, and the primary interaction theory instead favoring implicit, inter-subjective responsiveness and non-conceptual, interpersonal know-how—none of the theories in question has grounds to deny that we in fact produce and employ the full range of such predictive, action-guiding representations. GIM, therefore, takes no position on this issue, nor (as mentioned above) on the question of whether, given the diversity of representations under consideration, a single module could account for the production of all of them. GIM may well turn out, instead, to involve the cooperation of several distinct systems.[35]

All this being said, we would like to suggest that GIM might play a role not just in attributing intentional states to others, but also in attributing them to ourselves. That we might less often be in need of such representations in our own case is no argument against the suggestion that, when we do need them (e.g. when we are trying to predict the outcome of a multi-agent interaction in which we ourselves are among the agents) we turn to GIM; and, indeed, such a supposition is consistent with the general principles of functional specialization. Further, that we *would* sometimes need such self-interpreting seems quite likely when we consider that many intentional states are likely to be instantiated as distributed brain states, produced as often as not by phylogenetically ancient modular systems, and intimately tied to a set of physiological effects and near-reflexive responses (which is to say that they are both informational and dispositional in character, what Millikan has called pushmi-pullyu representations (Millikan, 1996), but which would not necessarily be available, as such, to higher-order cognitive processes. This underlines the fact that *being* in a given state need not be the same thing as *representing* one's self as being in that state, and, more importantly, that the state itself need not be instantiated in a form appropriate for certain kinds of cognitive operations. GIM might be able to bridge any such gaps.[36]

[35] Consider the distinctions made between systems of primary and secondary intersubjectivity in, e.g. (Gallagher, 2004; Trevarthan, 1979), and between the 'Eye Detection Detector' (EDD), the 'intentionality detector' (ID), the 'shared attention mechanism' (SAM) and the 'theory of mind' (ToM) modules in (Baron-Cohen, 1995).

[36] There is another class of objection to such an hypothesis, which argues that any account of introspection which relies on representations of mental states necessarily entails that we are thereby aware only of these representations, and not of the states themselves. This is a common confusion which ought to be easily addressed: as with any theory of representational awareness, this one suggests not that we are aware of *representations*, but that representations are (can be) the means by which we become aware of entities. As William Seager writes: "The key to understanding this position on introspection is always to bear in mind that when we perceive we do not perceive a perceptual state

In any event, the supposition that the same system is involved in interpreting others as well as ourselves is supported by studies of autism. For instance, the data presented in (Baron-Cohen 1989; 1991) are consistent with the prediction that damage to the systems that help us represent the intentional states of others would also impair self-interpretation; Baron-Cohen found that on false-belief tasks, autistic subjects have as much trouble attributing these beliefs to themselves as they do to others.[37] However these various issues are eventually resolved, what interests us here is that for GIM to play a role in self-interpretation requires only the assumptions that (1) GIM can make attributions to any object for which it is given input data, and (2) GIM utilizes and respects the object labeling and identification afforded by other modular processes. In the case of self-interpretation, intentional attributions need only be appropriately marked with the (generic) self-referential token "SR*". Further, as the basis for such self-specification is the somatoceptively grounded "I-context", which by its structure identifies one and only one object, one's own body, this appears to provide some explanation for the oft-observed fact that self-ascriptions of mental states are immune to error through misidentification of their subject. Of course, it is immediately obvious that giving this immunity organic-structural, which is to say fallible, basis opens up the very possibility of the kind of error that is meant to be impossible. If only one's "wires" get crossed or one's brain is damaged in the right way, it would appear that one could, indeed, be in the position of assigning to one's self mental states which were not one's own, or of assigning to another one's own mental states. Hogan and Martin (2001) discuss four such cases, including the case of a telepath who is not always sure of whose mental states he is aware.

but rather we perceive what the perceptual state represents. Seeing a tiger involves a representation of a tiger but it does not involve *seeing* that representation." (Seager, 2001 p.259)

[37] For a thorough defense of the claim that the facts of autism show that introspective awareness of mental states is a product of a theory of mind, see (Carruthers, 1996a). For an alternative interpretation of the data on autism, supporting instead the primary interaction view, see (Gallagher, 2004). Note that Gallagher's interpretation of autism does not undermine the narrow claim made above, that the data is compatible with the supposition that the same system could be involved both in the intentional interpretations of others and of ourselves.

> *Example 1.* John is telepathic. If through introspection he is aware of the occurrence of mental properties, then in the absence of contextual clues, he is in a position to infer only that someone has the properties, not that they are his own. John is also a hardened egoist. He has never had a sympathetic feeling for another person. In fact, because of how he is constituted psychologically he is incapable of having a sympathetic feeling. But he does not know that. On the occasion in question John discovers through introspection that someone is having a sympathetic feeling. The ordinarily reliable contextual clues are present, but this time they lead him astray. He judges that it is himself who is having the sympathetic feeling—actually it is someone else. (p.208)

The case of John is not unlike the thought experiment presented earlier, wherein the robot JP-B4 had access to the perceptual and belief states of KQ-C5, but did not bother to distinguish its own states from KQ-C5's.[38] In such a case JP-B4's introspection is not guaranteed to turn up only its own states. JP-B4's introspectively generated belief "I am seeing a tiger" (or, "JP-B4 is seeing a tiger") is mistaken just because it is, in fact, KQ-C5 and not JP-B4 who is seeing the tiger. Thus, one prediction of the hypothesis that introspective awareness is the result of something like GIM is that if somehow the identification and labeling of objects goes awry, so too will the ability to accurately introspect. Insofar as we accept the fact that we are organic beings, and our mental life must somehow depend upon our physical constitution, it is not clear how such possibilities can be denied. Naturally, when one *does* mis-attribute a mental state, it may thereby *seem* as if the state is one's own, and one may well act accordingly[39]—but this would be nothing but an effect of the operation of the system as a whole; to rest the guarantee against mis-identification on this would be to make it into a truism (whenever it seems as if one is in a certain mental state, it will seem as if that mental state is one's own). Of course it may be

[38] Hogan and Martin also discuss a machine version of their telepath case.

[39] This works most cleanly for factual beliefs which would ordinarily have mostly inferential roles, and little connection with the sort of visceral feelings that, for instance, a *fear* might have. If one misattributed to one's self the belief that the car keys are in the drawer, this belief, being labeled with one's self-representing token "SR*", would indeed seem to be, and would function as, one's own belief, and as a result one would indeed look in the drawer for the keys. However, it is less clear what the effect would be in the case of a misattribution of the fear of snakes. Presumably, believing this of one's self, one would try to avoid snakes; however, upon seeing a snake, if the usual reactions of fear were not present, one might well conclude that one was no longer afraid of snakes. As noted already above, it is no part of the function of GIM to *create* intentional states and their attendant effects; rather GIM

that such seeming was always, in fact, the basis of the guarantee. Thus, it is perhaps it worthwhile to reformulate this guarantee along the following lines: insofar as

(10) One is able to identify one's self, and

(11) One is able to attribute mental states to particular objects, and

(12) The mechanism whereby mental states are attributed to objects respects the identification and labeling of those objects,

then it is guaranteed that the self-attribution of mental states will be immune to error through mis-identification of the bearer of those states. Since, according to our hypothesis, self-identification is made possible by the grasp we have on our bodies in virtue of the structure of somatoception and its cooperation with other forms of perception, the attribution of mental states to objects is made possible by GIM, and the respect for object labeling is a function of the general and necessary cooperation between modules with different purposes (that must nevertheless make attributions of properties to the same object), this suggests three distinct areas where malfunctions would lead in part to trouble with introspection. John the telepath, for instance, could be supposed to fail on either criterion (11) or (12), depending on whether we interpret his defect as the inability, in detecting a mental state, to always attribute it to an object, or as a problem with maintaining the identity of the object in question when attributing mental states to it.[40]

It so far appears that, on the assumptions that self-identification is rooted in somatoception, and that this self-identification is respected by GIM (which is responsible for generating representations of mental states), we can account for some central aspects of self-representation:

generates representations of such states for the purpose of reasoning about them. There are some interesting empirical and theoretial issues here, but unfortunately we have no room to pursue them.

[40] The criteria (10)-(12) thus could be used to suggest a number of testable predictions regarding the kinds of defecits one might expect in the self-attribution of mental states.

its motivational force, the general immunity to error through mis-identification of the subject in the self-attribution of subjective properties, and the apparent differences in the character of our awareness of our own mental states and those of others.

So, the question is: how far does this get us? Does the self-representing mental token we have hypothesized, based on somatoceptive resources and shared with other mental modules, have *all* the properties generally associated with the first person? Clearly, we haven't the space for any such comprehensive recounting of, nor for the attempt to naturalize, the myriad properties supposed to be associated with first-person reference and representation. We can claim only to have laid a foundation for some such future account. Still, it is possible here to outline our approach to accounting for one final aspect of self-awareness: the self-reflexive character of some first-personal representations, those that represent the self as representing. Interestingly, this aspect of self-representation may be implicit in the operation GIM itself, at least in its more advanced and abstract manifestations. For one kind of situation with which GIM is specifically suited to deal is one in which an explanation of events involves (or even requires) the attribution of an intentional state to an agent with a content not warranted by (one's own assessment of) reality. Thus, we explain the dog barking up the tree in terms of his belief that the squirrel is up there, even though we have seen the squirrel leap away. Indeed, the general test for a child's ability to abstractly represent the intentional states of others is some version of a false-belief task. For instance, Ann watches Sally put her marble into a basket, and then leave the room, whereupon Ann moves the marble out of the basket into a box. When asked where Sally will look for her marble, a child who has not yet developed the ability to attribute beliefs to others will say "in the box",[41] while a child *with* the ability in question will say "in the basket" (Leslie, 2000).

[41] Recall that it is no part of the theory of mind hypothesis that a child too young for a theory of mind does not *have* beliefs, only that he or she cannot represent and reason with these beliefs.

Thus, implicit in the child's understanding and treatment of *abstract* representations provided by GIM is that they are *representations*, and thus can diverge from reality. When one represents *one's self* as being in a belief state, it appears to be likewise implicit that one is *representing* something, and it may not in fact be the case. Although the ability to treat representations in this way, at least explicitly, may well require capacities not here discussed,[42] it does not appear that any special form of *self-representation* is called for.

One final implication of this picture is worth bringing to the fore. In our explicit, conceptual representations of mental states, the self-other distinction is manifested in the syntactic attachment of mental-state concepts to subject tokens denoting one's self or another. Because GIM works by attaching intentional representations, and, where appropriate, mental-state concepts, to already identified objects, this implies that in *introspective reflection*, every mental state will be accompanied by the self-referring token. Consider, for instance, the experience of seeing a red bus. In the first instance, seeing the red bus does not require introspection; rather it involves a visually generated representation in virtue of which one is aware of the red bus. But *reflecting* that one is seeing the red bus *does* involve introspection, the operation of GIM, and the self-attribution of the mental state "seeing a red bus". The self-attribution takes the form of the predication of "is seeing a red bus" of the self-referring mental token. The same is true of the original, GIM-generated representation that Mary is sad. One can be aware of Mary's sadness, and respond appropriately to Mary, without reflecting on it. However, when one does reflect, what is involved is the attachment of this representation to the self-token: "I (SR*) believe(s) that Mary is sad". Although it is an interesting aspect of our (and, indeed, of any sufficiently expressive) representational system that indefinite regress is possible ("I believe that I believe that Mary is sad"), this needn't imply the inevitability of *infinite* regress in any specific case. One

[42] For instance, mastery of the concept of a belief, or of a mistake.

may remain at any given representational level, being thereby aware of one's own seeing of the red bus or of one's belief that Mary is sad; but one may also reflect on *this* representation, and introspect (for instance) that one believes one is seeing a red bus, or is now experiencing the seeing of the red bus, or some such. The suggestion is not that a higher level of representation is required in any given case, but only that, when introspection is asked to deliver another representation of mental contents, it is part of the normal functioning of GIM to tag the representation it delivers with the token marking its bearer. In introspective reflection, then, everything comes to awareness tagged with the self-referring mental token.[43]

5. Conclusion

We have shown, first of all, that a self-referring mental token need not be indexical. Instead, all that is required is for it to be connected in the right ways with the information-gathering and action-producing components of a system so as to make it self-referential according to criteria (7)-(9). We have further suggested that a mental token with the required properties—"SR*"—is a direct result of the basic function of the human somatoceptive and motor systems. Because of its specialized structure, the somatoceptive system has a firm grasp on the self—what we have called the essential prehension—in virtue of which it produces self-specifying representations with just the right content and connections to make an information bridge to, and allow the proper organization of, other information about the self. More importantly, "SR*", in virtue of its grounding in somatoception, has the right connections with our action-guiding systems to account for the special motivational properties of the information organized under it, and the apparent self-directedness of certain actions.

[43] (Perlis, 1997) contains the stronger claim that *every* conscious mental content is accompanied by a sense of the self. It may be that this stronger claim is also warranted by the theory presented here of the bodily grounding of self-

Finally, we have suggested that the special properties generally associated with first-person knowledge can be accounted for simply by allowing that the self-identification provided in virtue of the essential prehension is respected by other representation-generating mental modules, e.g. vision, natural language, GIM, etc., and that therefore the same self-representing token (same by whatever criterion of sameness applies in the realm of modular mental processes), is attached to representations produced by these modules. For instance, we have suggested that indexically expressed self-knowledge may be the result of the use of "SR*" by the language module, which expresses information represented with "SR*" by the use of "I", "Ich", "yo", "je" or whatever local convention dictates. We see no immediate obstacle to accounting for this translation in terms of the general necessity for the various modules of the mind to respect the identification of objects as they work together each in their specialized domain. What results from these various attachments are self-referential representations with all of the familiar properties of first-person tokens: the special motivational significance of the information predicated using it, the general immunity to error in identification of the subject in the case of the self-attribution of subjective properties, indexicality, and self-reflexiveness.

References

Anderson, M. L. 2003a. Embodied cognition: A field guide. *Artificial Intelligence*, 149(1): 91–130.

Anderson, M. L. 2003b. Representations, symbols and embodiment. *Artificial Intelligence*, 149(1): 151–6.

Anscombe, G. E. M. 1963. *Intention, 2ed.* Cambridge, MA: Harvard University Press.

Ballard, D. H., Hayhoe, M. M., Pook, P.K., and Rao, R.P.N. 1997. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4): 723–42.

Baron-Cohen, S. 1989. The autistic child's theory of mind: A case of specific developmental delay. *Journal of Child Psychology and Psychiatry*, 30: 285–97.

Baron-Cohen, S. 1991. Precursors to a theory of mind: understanding attention in others. In A. Whiten (ed). *Natural theories of mind: evolution, development, and simulation of everyday*

awareness, but we will leave this arguement to some future paper.

mindreading. (pp. 233–51). New York: Blackwell.

Baron-Cohen, S. 1994. How to build a baby that can read minds: Cognitive mechanisms in mindreading. *Current Psychology of Cognition*, 13: 513–52.

Baron-Cohen, S. 1995. *Mindblindness: an essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Bermúdez, J. L. 1998. *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Botvinick, M., and Cohen, J. 1998. Rubber hands 'feel' touch that eyes can see. *Nature*, 391: 756.

Brooks, R. 1991. Intelligence without representation. *Artificial Intelligence*, 47: 139–60.

Carrozzo, M., McIntyre, J., Zago, M., and Lacquaniti, F. 1999. Viewer-centered and body-centered frames of reference in direct visuomotor transformations. *Experimental Brain Research*, 129(2): 201–10.

Carruthers, P. 1996a. Autism as mind-blindness: an elaboration and partial defence. In P. Carruthers and P.K. Smith (eds). *Theories of Theories of Mind*. (pp. 257–73). Oxford: Basil Blackwell.

Carruthers, P. 1996b. Simulation and self-knowledge: a defence of theory-theory. In P. Carruthers and P.K. Smith (eds). *Theories of Theories of Mind*. (pp. 22–38). Oxford: Basil Blackwell.

Carruthers, P. 2002. The cognitive functions of language. *Behavioral and Brain Sciences*, 25(6).

Castañeda, H-N. 1966. 'He': a study in the logic of self-consciousness. *Ratio*, 8: 130–57.

Chalmers, D. 1995a. Absent qualia, fading qualia, dancing qualia. In Thomas Metzinger, editor, *Conscious Experience*. Imprint Academic.

Chalmers, D. 1995b. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2: 200–219.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. 1979. On cognitive structures and their development: A reply to piaget. In M. Piattelli-Palmarini (ed). *Language and Learning–The Debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Chrisley R. 2003. Embodied AI. *Artificial Intelligence*, 149(1): 131–50.

Craig, J. C., and Rollman, G. B. 1999. Somesthesis. *Annual Review of Psychology*, 50: 305–31.

Edelman, S. 2002. Constraints on the nature of the neural representations of the visual world. *Trends in Cognitive Science*, 6: 125–31.

Ehrsson, H. H. 2002. Awareness of limb movement: Illusory sensations in the primary motor cortex. *Proceedings of Tucson 2002: Towards a science of consciousness*, 65–6.

Fodor, J. 1983. *The Modularity of Mind*. Cambridge, MA: MIT Press.

Frege, G. 1960. On sense and reference. In P. Geach and M. Black (eds). *Translations from the Philosophical Writings of Gottlob Frege*, pages 56–78. Oxford: Basil Blackwell.

Gallagher, S. 1986. Body image and body schema: A conceptual clarification. *Journal of Mind and Behavior*, 7: 541–54.

Gallagher, S. 2001. The practice of mind: Theory, simulation, or interaction? *Journal of Consciousness Studies*, 5-7: 83-108.

Gallagher, S. 2003. Bodily self-awareness and object perception. *Theoria et Historia Scientarum: International Journal for Interdisciplinary Studies*, 7(1).

Gallagher, S. 2004. Understanding interpersonal problems in autism: Interaction theory as an alternative to theory of mind. *Philosophy, Psychiatry and Psychology*.

Gallistel, C. R. 1990. *The Organization of Learning*. Cambridge, MA: MIT Press.

Ghez, C., Gordon, J., and Ghilardi, M. F. 1995. Impairments of reaching movements in patients without proprioception. ii. effects of visual information on accuracy. *Journal of Neurophysiology*, 73: 361–72.

Goldman, A. I. 1989. Interpretation psychologized. *Mind and Language*, 4: 161-185.

Gordon, R. M. 1986. Folk psychology as simulation. *Mind and Language*, 1: 158-171.

Gordon, R. M. 1995. Simulation without introspection or inference from me to you. In M. Davies and T. Stone (eds). *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell Publishers.

Hari, R., Hanninen, R., Makinen, T., Jousmaki, V., Forss, N., Seppa, M., and Salonen, O. 1998. Three hands: fragmentation of human bodily awareness. *Neuroscience Letters*, 240: 131–134.

Heal, J. 1986. Replication and runctionalism. In J. Butterfield (ed). *Language, Mind, and Logic*. Cambridge: Cambridge University Press.

Heal, J. 1998a. Co-cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind and Language*, 13: 477-98.

Heal, J. 1998b. Understanding other minds from the inside. In A. O'Hear (ed). *Current Issues in Philosophy of Mind*. Cambridge: Cambridge University Press.

Hogan, M. and Martin, R. 2001. Introspective misidentification. In A. Brook and R. DeVidi (eds). *Self-reference and Self-awareness*. (pp. 205–13). John Benjamins Publishing Company.

Humphrey, N. 1992. *A History of the Mind*. New York: Simon and Schuster.

Hurford, J. R. 2003. The neural basis of predicate argument structure. *Behavioral and Brain Sciences*, 23(6).

Kirsh, D. 1991. Today the earwig, tomorrow man? *Artificial Intelligence*, 47(3): 161–184.

Leslie, A. M. 1994. ToMM, ToBy and Agency: Core architecture and domain specificity. In L. A. Hirschfeld and S. A. Gelman (eds). *Mapping the mind: Domain specificity in cognition and culture*. (pp 119–48). Cambridge: Cambridge University Press.

Leslie, A. M. 2000. How to acquire a ‘representational theory of mind’. In D. Sperber (ed). *Metarepresentations: A multidisiplinary perspective*. (pp. 197–223). Ablex Publishing.

Marr, D. 1982. *Vision: a computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman.

Millikan, R. G. 1990. The myth of the essential indexical. *Noûs*, 24: 723–34.

Millikan, R. G. 1996. Pushmi-pullyu representations. In L. May and M. Friedman (eds). *Mind and Morals* (pp.145–61.) Cambrddge, MA: MIT Press.

Milner, A. D., and Goodale, M. 1995. *The Visual Brain in Action*. Oxford: Oxford University Press.

Nicolelis, M., Ghazanfar, A., Stambaugh, C., Oliveira, L., Laubach, M., Chapin, J., Nelson, R., and Kaas J. 1998. Simultaneous encoding of tactile information by three primate cortical aread. *Nature Neuroscience*, 1(7): 621–630.

O’Donovan-Anderson, M. 1997. *Content and Comportment: On Embodiment and the Epistemic Availability of the World*. Lanham, MD: Rowman and Littlefield.

O’Shaugnessy, B. 1980. *The Will: A dual aspect theory*. Cambridge: Cambridge University Press.

Perlis, D. 1997. Consciousness as self-function. *Journal of Consciousness Studies*.

Perlis, D. 2000. The role(s) of belief in AI. In J. Minker (ed). *Logic-based AI*, chapter 14.

Perry, J. 1977. The problem of the essential indexical. *Noûs*, 13: 3–21. Reprinted in A. Brook and R. DeVidi *Self-reference and Self-awareness*. (pp. 143–59). John Benjamin Publishers, 2001.

Perry, J. 1990. Self-notions. *Logos* 1990: 17-31.

Perry, J. 1993. *The Problem of the Essential Indexical and Other Essays*. New York: Oxford University Press.

Perry, J. 1995. Rip van Winkle and other characters. *The European Review of Analytical Philosophy* Volume 2: *Cognitive Dynamics*, 13-39.

Perry, J. 1998. Myself and "I". In M. Stamm (ed). *Philosophie in Synthetischer Absicht*. (pp. 83-103). Stuttgart: Klett-Cotta.

Pinker, S. 1994. *The Language Instinct*. New York: Harper Perennial.

Rosenberg, G. and Anderson, M. L. 1994. A brief introduction to the guidance theory of representation.

Rosenberg, G. and Anderson, M. L. forthcoming. Content and action: The guidance theory of representation.

Scholl, B. J. and Leslie, A. M. 1999. Modularity, development and 'theory of mind'. *Mind and Language*, 14(1): 131–53.

Seager, W. 2001. The constructed and the secret self. In A. Brook and R. DeVidi (eds). *Self-reference and Self-awareness*. (pp. 247–68). John Benjamins Publishing Company.

Shoemaker, S. 1968. Self-reference and self-awareness. *Journal of Philosophy* 65: 555-67.

Spelke, E.S. 1990. Principles of object perception. *Cognitive Science*, 14: 29–56.

Trevarthen, C. 1979. Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (ed). *Before Speech*. Cambridge: Cambridge University Press.