Ch. Kanzian, J. Quitterer, E. Runggaldier (Hg.)

Persons. An Interdisciplinary Approach.

Wien: öbv&hpt 2003

Self-Consciousness in Cognitive Systems[1]

ANSGAR BECKERMANN, BIELEFELD

1.

Putting it a little provocatively, the question I want to address can be phrased thus: "Which cognitive systems have a self?" Or: "Can artificial cognitive systems ever have a self?" Talking about *a* or *the* self, however, seems to me to be at least dubious. First of all, the word 'self' is usually used as a pronoun which occurs in sentences like "He is coming himself" or "I don't know whether our chancellor will come himself". The noun 'self', in contrast, typically occurs in certain idioms as 'my humble self' or 'your good selves'. The apotheosis of this noun occurring in sentences like "Each person's self is indestructible", however, is simply a philosophical fabrication, probably due to John Locke. At least in the second book of his *Essay Concerning Human Understanding*, Locke writes:

> Self is that conscious thinking thing, (whatever Substance made up of whether Spiritual or Material, Simple or Compounded, it matters not), which is sensible, or conscious of Pleasure and Pain, capable of Happiness or Misery, and so is concern'd for it *self*, as far as that consciousness extends. (ECHU II xxvii 17)

The linguistic background of this rather peculiar use of the word 'self' may have been that in Locke's time words like 'myself' and 'itself' were written apart. This, of course, may have led to the assumption that there might be such a thing as my self. It should be noted, however, that in Latin it is not even possible to use 'ipse' as a noun. And, therefore, it is more than remarkable that in a recent English edition of Descartes' *Meditations* we find as a translation of the rather innocuous sentence "Nunquid *me ipsum* non tantum multo verius, multo certius, sed etiam multo distinctius evidentiusque cognosco?" (Descartes *Meditationes*, 33 – italics mine) the English phrase: "Surely my awareness of *my own self* is not merely much truer and more certain … but also much more distinct and evident."[2]

But even if the new and, in a way, exceptional use of the word 'self' by Locke may strike one as annoying, it is not necessarily harmful. For Locke makes quite clear how he wants the term 'self' to be understood: a self is that thing which is capable of undergoing pleasant and painful sensations, which may experience happiness or misery, and which, therefore, is concerned for itself. Thus the self seems to be nothing other than a human being or a person.[3] At least, Locke himself deliberately leaves open the question whether the self is something material or spiritual, i.e. immaterial. He does not want to take a stand with regard to Cartesian Dualism, but he seems at least to have acknowledged the possibility that Descartes might be right on this issue, i.e., that the real self is a *res cogitans*. Maybe this is why talk of 'selves' is often understood as having Cartesian connotations.

Be that as it may, a self, according to Locke, simply is that which is capable of having pleasant and painful sensations and which, therefore, may experience happiness or misery. In the recent literature, however, another feature has become closely associated with the concept of a self. Lowe, e.g., writes:

> [Selves] are subjects of experience which have the capacity to recognise themselves as being individual subjects of experience. Selves possess reflexive self-knowledge. By 'reflexive self-knowledge' I mean, roughly speaking, knowledge of one's own identity and conscious mental states – knowledge of who one is and of what one is thinking and feeling. … [R]oughly speaking – having the kind of reflexive self-knowledge which makes one a person goes hand-in-hand with possessing a 'first-person' concept of oneself, the linguistic reflection of which resides in an ability to use the word 'I' comprehendingly to refer to oneself. (Lowe 2000, 264f.)

Apart from having the capacity to undergo experiences, selves, according to Lowe, are characterized by their possessing reflexive self-knowledge – i.e. knowledge the content of which the system itself can express only by using the word 'I'. But if that is what a self amounts to, one should refrain from asking questions such as "Which cognitive systems have a self?". For, first, this question really should be put thus: "Which cognitive systems are selves?". And, second, this question can very well be expressed in perfectly ordinary English: "Which cognitive systems are capable of reflexive self-knowledge?". This question – which simply does not insinuate the existence of mysterious entities labeled 'selves' – is the question I want to address in this paper.

2.

Cognitive systems are systems which represent the world they live in in order to cope with their environment. To keep things simple, I shall henceforth consider only systems whose knowledge of their environment is represented by way of lists, i.e. systems with a special kind of Language of Mind. What I want to argue is that such systems possess the reflexive self-knowledge necessary for self-consciousness if and only if they employ a special kind of representation, *viz.* representations which are about themselves and which furthermore play a very specific role in their overall cognitive architecture. I will not be concerned with any other cognitive system, yet I am quite sure that the considerations to follow can, more or less smoothly, be transferred to these systems.[4]

In his paper "Myself and I", John Perry draws some very helpful distinctions among different kinds of self-knowledge – *agent-relative knowledge*, *self-attached knowledge* and *knowl-*

---

2. Descartes, René *Selected Philosophical Writings* (86 – italics mine). In the 1911 edition by Haldane and Ross it still runs : "[D]o I not know *myself*, not only with much more truth and certainty, but also with much more distinctness and clearness?" (156 – italics mine).
3. Cf. Disraeli's remark: "Self is the only *person* whom we know nothing about". *The Oxford English Dictionary*, 2nd edition, s.v. 'self', C.I.1.e. (italics mine).

4. The story by means of which I am going to answer this question is, of course, not entirely new. Cf. e.g. Rosenberg 1986, esp. ch. VI and VII.

edge of the person one happens to be. Agent-relative knowledge is knowledge of the environment that is represented from the perspective of a particular agent. Knowledge of this kind does not presuppose any representations which refer directly to the agent himself. As Perry puts it: "[T]he agent need not have an idea of self, or a notion of himself or herself." (Perry 1998, 83) In other words: In the lists used to represent agent-relative knowledge there need not occur *any one symbol* that refers to the agent himself. In the end, agent-relative knowledge consists in representations already well understood by cognitive scientists – representations in which the environment is represented not by means of world coordinates (Cartesian, external coordinates) but by means of what is called an ego-centric reference system (body reference system). Let's look at an example.

One of the early successes of AI was the program SHRDLU. What its designer Terry Winograd aimed at was to combine the capacities of language processing and problem solving within one system. SHRDLU 'lives', as it were, in a micro-world containing various kinds of objects – blocks, spheres, and pyramids as well as a box. Its tasks mainly consist in arranging these objects in new ways – e.g., to put the green pyramid on top of the red block, or the sphere into the box. For my purposes, however, the only aspect of real importance is the way in which SHRDLU represents the micro-world it lives in.
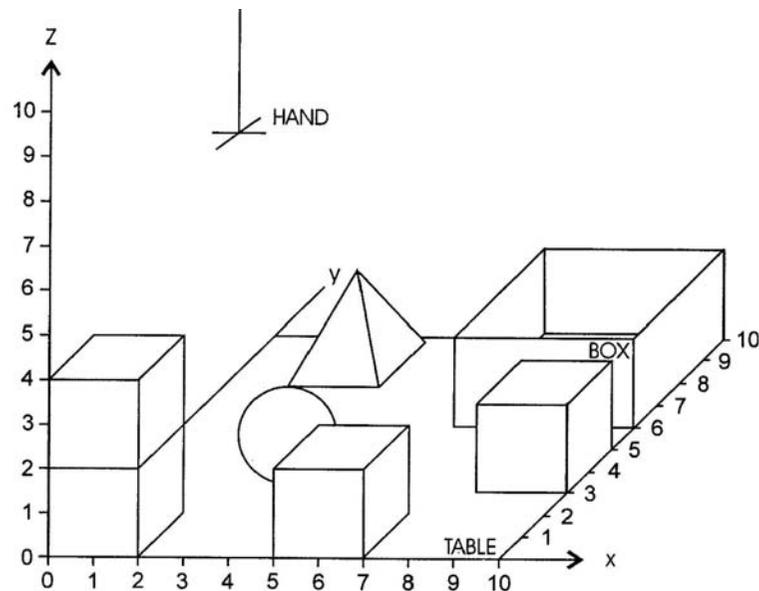

Figure 1

A situation like that shown in fig. 1, e.g., is typically represented in the form of lists like the following:

```
(is-a      object-1    block)
(color     object-1    green)
(place     object-1    (1 1 2))
(size      object-1    (2 2 2))
…
(is-a      object-5    sphere)
(color     object-5    green)
(place     object-5    (4 3 0))
(size      object-5    (2 2 2))
…
(grasping  hand        nil)
(place     hand        (2 5 7))
```

What's decisive here is that all information concerning the *place* of the objects in question is represented in world coordinates. And that this is true even for the hand of the system itself. If SHRDLU is going to grasp a certain object *a* it, therefore, has to find out: Where is the top of *a*? Where is the hand? How can the hand move from its place to the top of *a*? All this seems to be incredibly unnatural. Obviously, we ourselves do not represent the whereabouts of the objects in our vicinity in terms of objective world coordinates. And, as a matter of course, we certainly do not represent our own place in the environment in this way. Looking from the door into my study I rather naturally represent what I see in the following way:

> There is a desk about 5 steps in front.
> There is a window immediately behind the desk.
> On the left side on top of the desk there is a monitor.
> Immediately in front of the monitor there is a keyboard.
> There is a cup in the middle of the top of the desk.
> There are bookshelves on the left side of the desk.
> At half height in one of the shelves there is a printer.

That is, I represent the locations of the objects in my environment by means of the spatial relations in which these objects stand *to me* and *to each other*. There are at least two very good reasons for doing it this way. With regard to the first reason John Perry writes:

> Everything we learn about other objects we learn by employing methods that are appropriate because those objects stand in certain relations to us. … [The objects in our vicinities have] *agent-relative roles*: roles that other individuals play in the lives of agents. These are agent-relative roles, because an object plays or doesn't play such a role relative to a given agent, at a given time. For example, my computer is playing the role of *object in front* right now, relative to me, but not relative to you. … This is the first of two very general facts I want to emphasize: any object we learn about plays some agent-relative role, basic or derived, in our life. We learn about the object by using an epistemic method connected to the

role, a way of finding out about the object or person playing that role. The way to find out about the object in front of you is to look at it, or perhaps to walk up to it and touch it. (Perry 1998, 84f.)

In other words, all we know about the objects in our environment we know because these objects stand in certain relations to us, because they play certain agent-relative roles with regard to us. It's no wonder, therefore, that we represent the objects we encounter as bearers of just these roles – as the table five steps in front of me, the person to my left, the floor under my feet, etc. However, there is another point that is even more important. Only if I represent the objects around me as bearers of the roles they play with regard to me am I able to know immediately how what is going on in my neighborhood will affect me. Think of a ball that is just about to fly right into my face. If I represent this ball as an object which is going to hit me in the next second it's immediately obvious that I have to do something – to catch the ball, to duck or what have you. If, however, I represent the movement of the ball in terms of world coordinates, I don't know what this means for me unless I also know – again in terms of world coordinates – where I am, and unless I figure out whether the path of the ball and my own path through the world cross at a certain point.

It is thus quite obvious that agent-relative representations make it much easier for us to assess the relevance of what is going on in the environment for our own concerns. This leads to the second reason Perry states:

> … [E]verything we do comes down to performing operations on the objects around us – objects in front of us, behind us, above us; objects we are holding; objects we can see. By doing these things, we do things to objects in less basic relations to us. … I know how to move my body so as to effect objects around me, and I know how effecting those objects will effect other objects related to them in certain ways. (Perry 1998, 85)

If we want to do something we always have to carry out bodily movements that have an effect on the objects around us in virtue of the relations these objects bear to us, the agents. There is a certain *type* of bodily movement by which we grasp a cup on a table immediately in front of us. There is a certain type of bodily movement by which we scare off a fly on our left hand. And there is, to cite Perry's favorite example, a certain type of bodily movement by which we stab the person standing to the left of us. But there is no type of bodily movement by which we put the pyramid at (a b c) on top of the block at (d e f). That is to say, if we try to figure out by which kind of movement we can perform a certain action, it is tremendously helpful to represent the objects in our environment by way of the agent-relative roles they bear to us. That is because there is a type of movement by which I can grasp a cup on a table immediately in front of me and because I know this – while, on the other hand, I have no idea by means of which bodily movement I could grasp a cup being located at place (7 10 3).

> There are then two kinds of methods connected with agent-relative roles, epistemic methods and pragmatic methods. These two kinds of methods are the key to all human intelligence and purposive activity. We know how to find out what kinds of objects occupy these roles, and we know how to perform various operations on them. … Our practical knowledge then, the knowledge that enables us to do things, forms a structure at whose

base is information about the objects that play relatively basic agent-relative roles in our lives. (Perry 1998, 85)

How does all this relate to the topic of self-consciousness? Well, at least at first sight, it seems as though agent-relative representations must always involve an explicit reference to the agent. There appear to be reasons why cognitive scientists use the term '*ego*-centric reference system'. If, for example, I see an apple before me, pick it up and eat it, the complex movements that result in getting the apple into my mouth are based on what I learned from perception, that is the distance and direction of the apple from *me*. If I am now using the keyboard of my computer, "I have to move my fingers a certain distance and direction from *me*. It isn't enough to know where the buttons were relative to one another, or where the [keyboard] was in the building or the room. I had to know where these things were relative to *me*. It seems then, that these basic methods already require me to have some notion of myself." (Perry 1998, 86)

Thus, agent-relative representations seem to belong to those very representations the content of which can only be expressed by using the word 'I'.[5] But, actually, this is not true. As Perry himself points out, if we are to express the content of these representations by means of, say, English sentences, we always can do without personal pronouns like 'I'. If I want to represent a situation in which there is an apple a little bit to the right half a meter in front of me, I don't have to explicitly represent that the apple bears a certain *relation* to *me* – a relation the apple could also bear to other objects. What I have to represent is only that the apple has a certain property – the property of standing in a certain spatial relation-to-me. The second relatum, me, always remains fixed. Thus I don't need to keep track of it. In a way, it's the same as if I were to represent the situation in which there is an apple a little bit to the right half a meter in front of me by means of the polar coordinates r and α:

$$(\text{is-a} \quad \text{object-a} \quad \text{apple})$$
$$(\text{place} \quad \text{object-a} \quad (0.5 \; 8°))$$

It's obvious that such a representation does not involve any expression that explicitly refers to me. As I have already said, we can do entirely without such a reference. The general point is this:

> Sometimes all of the facts we deal with involving a certain *n*-ary relation involve the same object occupying one of the argument roles. In that case, we don't need to worry about that argument role; we don't need to keep track of its occupant, because it never changes. (Perry 1998, 87)

Agent-relative representations, thus, comprise a certain kind of self-knowledge: they comprise knowledge of how the objects in the environment are related to the agent. And, therefore, they also comprise knowledge about the agent himself. But these representations do not presuppose that the agent has a notion of himself. They need not contain any expression explicitly referring to the agent. And this, of course, is true also of the representations which comprise knowledge about the internal states of the agent. Sensations, for example, may very well be represented in the following way:

---

5. To be more precise: by using what Castañeda calls 'quasi-indicators'.

A pain in the left knee.
A tingling sensation in the stomach.
Left arm itching.

That is, even with regard to representations of the internal states of the agent, there is no need to use any expression that explicitly refers to the agent at all. The reason for this is simple. The cognitive systems we are looking at at this stage of the story don't represent other systems as having pains in their left knees or tingling sensations in their stomachs. At this stage, sensations and other internal states are always sensations and internal states of the cognitive system itself. Therefore, reference to the system can once again be dropped.

It is only when we consider systems possessing, as it were, genuine self-knowledge (self-attached knowledge, as Perry puts it) that we will encounter systems bound to have explicit self-notions. Thus, the next questions I want to address are: What kind of representations are needed for genuine self-knowledge? And how can such representations evolve?

3.

Let me rehearse the basic idea: Cognitive systems are systems which represent the world they live in in order to cope with their environment. How does that work in detail? What are the necessary preconditions for a cognitive system – let us call him 'AL' – to be able to represent his environment? Which processes are at work here?

Obviously, the process of representing the world must begin with the causal traces the environment leaves on AL. What AL has to do is to reconstruct the environment from these traces. That is to say, by analyzing the traces AL has to manage to answer the following questions:

1. How many objects make up the present scene?
2. What kinds do these objects belong to?
3. Where are these objects?
4. Which properties do the objects have, and which relations do they bear to each other?

If AL manages to answer these questions, he may – and let's assume he will – go on to construct representations of his environment in the following way:

- Exploiting the answer to the first question, AL will give internal names to the objects in his environment – e.g., names like 'object-36', 'object-37', etc.

- Having answered the second question for each object encountered, AL will add to his system of representations a new list of the form

    (is-a object-*x* *type*).

- Exploiting the answer to the third question for each object, AL will add to his stock of representations a new list of the form

    (place object-*x* *coordinates*).

- Finally, having answered the fourth question, AL will add to his former representations new lists like

    (color   object-*x*   *color*)
    (size   object-*x*   *size*)
    (*relation*   object-*x*   object-*y*)
    a.s.o.

That's the basic idea. But there are many cases in which that's not enough. For cognitive systems also need to solve the problem of recognizing objects they have encountered in previous scenes. If another stork approaches the nest, the stork in the nest not only has to find an answer to the question: Is it male or female? What is much more important is to find out whether the newcomer is its partner or an unfamiliar stork. If in analyzing a scene AL gives a *new* name to each object he encounters this in effect amounts to treating all objects as objects he never encountered before, and this, of course, could lead to entirely inappropriate reactions. That is, AL must – e.g., by comparing typical characteristic features – find out which of the objects in the currently examined scene are identical with objects he encountered before and for which he, therefore, already has internal names. With regard to all these objects AL has to replace the new internal name by the one already in use.[6]

For our purposes, it is crucial to note that none of this requires that AL employs a representation that explicitly refers to himself – i.e., an internal name for himself. He may very well represent the whereabouts of the objects he encounters in agent-relative coordinates, and he may build up representations indicating which relations-to-him these objects bear. As we have already seen, all this can be done without the use of any symbol that explicitly refers to AL. Thus, the question now is: under what conditions does it become, as it were, unavoidable for AL to use an internal name for himself? Let's proceed to tell the story step by step.

In the environment of cognitive systems, there are often other cognitive systems. And I shall assume that this is also true of the environment AL lives in. Now the most characteristic trait of cognitive systems in general is that their behavior not only depends on their, as you might want to call them, 'natural' properties, but also on how they represent the world. If AL is trying to predict what these strange objects in his environment are going to do, therefore, he is forced to develop a new kind of representation, *viz.* representations that tell him how the cognitive systems he encounters represent their shared world, which representations *they* build up in order to get along. Representations of this kind are commonly called 'meta-representations', and their general form is:

    (believes   object-*x*   *representation*)

or

    (wants   object-*x*   *representation*)

etc.

---

6. In his considerations on these issues Perry does not assume that the objects encountered get internal names. His idea is that all information concerning one object is stored in a particular file. On this account the present problem is this: With regard to each object AL encounters in analyzing a scene he has to decide whether to start a new file or to store the information received in an already existing file. (Cf. Perry 1998, 89ff.)

It is an interesting task to figure out what the representations that can be substituted for the variable '*representation*' might look like in detail. Sometimes AL may come to the conviction that a fellow cognitive system – say the system with the internal name 'object-111' – has a certain belief about an object AL himself is acquainted with and for which he uses the internal name 'object-7'. In this case, AL's meta-representation will look like this:

> (believes   object-111   (color object-7 green)).

By way of this meta-representation AL does, as it were, attribute a *de re* belief to his fellow cognitive system. For the content of this meta-representation is: object-111 believes of that object which AL (internally) calls 'object-7' that it is green.

But it also may be that the fellow cognitive system stares at something in front of its feet that AL cannot see, and that furthermore the system behaves in a way it behaves only in case it encounters a spider. How is AL going to represent this scene? Well, first, AL does not see the object in front of the fellow cognitive system. He, therefore, has to use a new internal name for this object – say, the name 'object-57'. Second, all that AL knows with respect to object-57 is that it is in front of the fellow cognitive system and that, obviously, this fellow cognitive system believes it to be a spider. Thus, AL will add to his system of representations the following two new representations:

> (in-front-of   object-57   object-111)
> (believes   object-111   (is-a   object-57   spider)).

Of course, it also may be that AL endorses the fellow cognitive system's belief and devises the representation

> (is-a   object-57   spider).

But that is of no importance to us. What is of importance is rather: AL has to use *new* internal names if he is to represent his fellow cognitive systems' representations dealing with things AL himself is not acquainted with.

Another thing is crucial here. It seems reasonable to assume that AL tries not only to represent the representations of his fellow cognitive systems but also their sensations. That is, AL will develop representations such as:

> (pain-in-the-knee   object-111).

For pains, pleasure and all the other sensations also have an effect on the behavior of AL's fellow beings.

However, the most important step is another one. Sooner or later AL will realize that his fellow cognitive systems represent him in the same way he represents them. That is to say, his fellows represent their environment as being inhabited by an object that is itself a cognitive system and that happens to be nobody other than – AL. Sooner or later, AL will notice that one of his fellows is staring at HIM or is approaching HIM in order to get something from HIM, e.g., food. Moreover, sooner or later it will no longer be feasible for AL to represent such situations

by means of an ego-centric reference system, i.e., in an agent-relative way. In order to represent the representations of his fellow cognitive systems that refer to him, AL will have to use an internal name – say, 'object-100' – for himself, something he did not do until that very moment. By means of this name he can then represent the wish of his fellow in the following way:

> (wants object-111   (gives-food object-100 object-111)).

Something similar will happen if AL notices that the fellow for which AL uses the internal name 'object-111' apparently comes to believe that he, AL, has a pain in his knee. For this belief can be represented by AL only like this:

> (believes object-111   (pain-in-the-knee object-100)).

There is no way of representing such a belief by using an ego-centric reference system. That is, AL can represent this belief of his fellow cognitive system only by using a new internal name that happens to be a name for himself. Yet sooner or later, AL will not only use this new name, he will also begin to realize that this name really is a name for HIMSELF. Admittedly, this is a metaphorical way of putting things. For how can AL realize that 'object-100' is a name for himself as long as he has no notion of himself? So let us try to spell out this metaphor. We have to address the question of what kind of process can count as the process of AL's developing a notion of himself. Three steps seem to be crucial.

1. AL starts to represent the representations of his fellow cognitive systems about him, and in doing so he begins to use a new internal name which happens to be a name for himself – AL.
2. AL starts to use this internal name also in representing situations where, e.g., he sees himself in a mirror.
3. A systematic connection evolves between representations containing the new internal name and those well-worn agent-relative representations which refer to AL only in a tacit way.

This last step seems to be decisive. That is, it is of utmost importance if in AL's cognitive architecture the representation

> (sitting-on object-3)

begins to play the same role as the representation

> (sits-on object-100   object-3),

and if the representation

> (pain-in-the-knee)

starts to play the same role as the representation

(pain-in-the-knee   object-100).

The effect of this process will be that in AL's cognitive architecture, older agent-relative representations become equivalent to representations explicitly referring to AL by means of the new internal name. Moreover, all the input that AL's cognitive architecture receives from AL's body, everything that is available by way of proprioception, will now yield representations that explicitly refer to AL. That is, AL will develop a body schema. So far, AL has represented his environment by 'asking': Which relation-to-me does object x bear? And he did so because the second relatum – AL himself – always remained fixed. Now he undoes this kind of rigidification because it turned out that he is just one out of many objects that may stand in the same relation to object x. Metaphorically speaking, AL begins to see himself through the eyes of others.

This has yet another important effect. AL begins to develop meta-representations with regard to himself. So far it simply was not necessary for him to know what he himself believed and wanted. Yet once he begins to see himself through the eyes of others, this is no longer so. For the behavior of his fellows also depends on what they believe about how AL represents the world. Hence, explicit knowledge of his own representations is no longer irrelevant to AL. Therefore, he had better begin to take them into account. Thus, finally, we arrive at what seems to be genuine self-knowledge. For this was characterized by Lowe in the following way:

> [By] 'reflexive self-knowledge' I mean, roughly speaking, knowledge of one's own identity and conscious mental states – knowledge of who one is and of what one is thinking and feeling. (Lowe 2000, 264f.)

4.

However, is this really genuine self-knowledge? Isn't the internal name 'object-100' an internal name like any other? How does it come about that representations containing this name have such a special status? Why do these representations in particular count as genuine self-knowledge? Why are these representations in particular responsible for AL's being self-conscious in the sense we are discussing here?

Well, first, it seems at least imaginable that AL together with his fellow beings begins to develop a language in which indexicals like 'I', 'you', 'there' etc. play their usual roles. (Perhaps this is even necessary for AL to become able to develop the representations and meta-representations already mentioned.) And what is also imaginable is that AL learns to express only those representations in which the internal name 'object-100' occurs by sentences containing the word 'I'. But wouldn't that be purely accidental? Wouldn't it have been equally possible for AL to learn to express only those representations in which the internal name 'object-13' occurs by sentences containing the word 'I'? The answer is No. For learning how to use the word 'I' implies learning that any member of the language community may use this word to refer to himself. Thus, AL will have learned the meaning of 'I' only if he has learned to use this word to express nothing but representations which are about himself. This is not enough, though. For this is the point where Perry's distinction between *self-attached knowledge*

and *knowledge of the person one happens to be* becomes relevant. Let us quickly rehearse just one famous example.

At the very beginning of "The Problem of the Essential Indexical" Perry writes:

> I once followed a trail of sugar on a supermarket floor, pushing my cart down the aisle on one side of a tall counter and back the aisle on the other, seeking the shopper with the torn sack to tell him he was making a mess. With each trip around the counter, the trail became thicker. … Finally it dawned on me. I was the shopper I wanted to catch. (Perry 1979, 33)

The point of the case is obvious. At first, Perry had the belief that a certain person – about whom he knew almost nothing – had a torn sack of sugar in her or his cart. After a while, however, Perry noticed that the person he was looking for was none other than himself. This way he came to believe that he himself had a torn sack of sugar in his cart. Looked at in a Russellian way, the two beliefs have the same content – the singular proposition <having a torn sack of sugar in the cart, John Perry>. However, the beliefs must be different since their consequences differ quite a bit. The first belief made Perry seek for the person with the torn sack in the cart to tell her or him about the mess she or he was making. The second belief, however, induced a different behavior. "I stopped following the trail around the counter and rearranged the torn sack in my cart." (Perry 1972, 33) Since the two beliefs have the same Russellian contents, they can differ only in their Fregean contents, i.e., in the modes of presentation involved. What is more, the Fregean content of the second belief seems to be very special since it involves a very special way in which Perry is given to himself – a way which might be dubbed the 'EGO-mode of presentation'. Thus, the question we have to address is this: What does the internal name 'object-100' that AL uses for himself have to do with such an EGO-mode of presentation?

One thing that is important here is that in AL's system of representations, different modes of presentation are usually accounted for by the use of different names. Suppose AL sees in a mirror that there is a bear behind some cognitive system, not realizing that the cognitive system he sees is actually himself. In this case, in representing the situation AL has to use a new name, say 'object-213', for the cognitive system seen. (Of course, the same holds for the bear unless AL is already acquainted with it.) In the end, this will lead to representations like

(is-a   object-511 bear)
(is-a   object-213 cognitive system)
(behind   object-511 object-213)

Only when AL comes to belief that the bear is behind his *own* back he will use the internal name 'object-100'.

Second, and this is even more important, different modes of presentation correspond to different ways of processing, or different functional/computational roles. If two internal names are computationally equivalent with respect to AL's cognitive makeup, they correspond to the same mode of presentation. On the other hand, two internal names which play different computational roles correspond to different modes of presentation. Thus, our question becomes: What special features of the computational role of the internal name 'object-100' make it suitable for corresponding to an EGO-mode of presentation?

Well, the way in which the name 'object-100' is processed in AL's cognitive architecture is very special indeed. Remember that we have assumed that only those representations of AL in which this name occurs are equivalent to corresponding agent-relative representations – 'being equivalent' here meaning simply 'having the same computational role'. This has two very important consequences. First, the entire proprioceptive input to AL's cognitive system yields only agent-relative representations and, by way of being equivalent, representations containing the name 'object-100'. Even if 'object-213' is, unbeknownst to AL, another internal name actually referring to AL, in AL's cognitive makeup this name will never be used to store information about his headaches or about the position of his feet – at any rate, it won't do so if this information is not coming from the outer sense organs, but from the internal proprioceptive system. Thus, the special way in which the states of AL's body are given to AL is connected only to the internal name 'object-100' and to no other name of that kind.

The second important consequence is this. Agent-relative representations have a characteristic feature – they immediately induce certain actions. Remember the agent-relative representation with the content "There is a ball just about to fly right into my face". At least in most cases, this representation will immediately lead to action. I'll duck, try to catch the ball or what have you. With a representation having the content "There is a ball moving with velocity v from place a to place b" things will be very different. This representation, taken in isolation, will probably have no effect at all on my behavior. Only if I realize that my own position is directly between a and b, will it likely cause me to do something. And there was another characteristic feature of agent-relative representations. If an agent represents the objects in his environment by means of the relations these objects bear to *him*, for a vast number of actions there will be a specific type of bodily movement by which the agent can carry out the action. If a bear is behind my back, I have to turn and, maybe, try to stab it. Being equivalent, representations containing the name 'object-100' inherit these two features from the corresponding agent-relative representations. That is, representations containing this name have the same immediate impact on action, and they are in the same way connected with certain types of bodily movements by which AL can perform certain types of actions.

This is decisive since *de se* beliefs also are characterized by the specific causal role they play within the cognitive system of a person and with regard to her or his actions. At least generally, they lead to typical egocentric reactions. Again, think of Perry's sugar example. Before noticing that he himself caused the mess, Perry's belief made him look for the person with the torn sack in the cart to tell her or him about the mess she or he was making. After that his reaction was quite different. He stopped following the trail and tried to rearrange the torn sack in his cart. Or take the strange experience once reported by Ernst Mach:

> Not long ago, after a trying railway journey by night, when I was very tired, I got into an Omnibus, just as another man appeared at the other end. "What a shabby pedagogue that is, that has just entered," thought I. (Perry 1998, 93)

This thought of Mach's may have led to nothing more than contempt or compassion on his side. But then Mach suddenly noticed that the man at the other end was none other than he himself.

> It was myself: opposite to me hung a large mirror. The physiognomy of my class, accordingly, was better known to me than my own. (Perry 1998, 93)

This caused a big shift. After having noticed that he himself was the man that seemed to be over there Mach evidently developed the thought "I look like a shabby pedagogue". And this thought certainly had entirely different consequences –shame, perhaps, or an attempt to clean his clothes.

Thus, in the end, we get an answer to our question "What special features of the internal name 'object-100' make it suitable for corresponding to an EGO-mode of presentation?" The answer is this: Representations containing this name play exactly the causal role that is characteristic of *de se* attitudes, i.e., of attitudes the content of which can only be expressed by using the word 'I'.[7]

<div align="center">5.</div>

Cognitive systems try to acquire knowledge of the world they live in because this considerably enhances the likelihood of their overall success. But in many cases successful action not only presupposes knowledge of the world, but also knowledge of the system itself. A cognitive system needs to know where *it itself* is located in the environment, whether *it itself* is threatened, which of *its* bodily organs are functioning, what needs *it* has, etc. In most cases, however, it will suffice to store this kind of self-knowledge in agent-relative representations – representations which, in a way, are about the system itself, but which do not contain any kind of explicit self-reference. Representations containing expressions that explicitly refer to the system itself are indispensable only if the cognitive system lives in a world inhabited by other cognitive systems. For in this case it will be necessary to represent these fellow systems as beings that themselves represent their environment – and, what is more, as beings that have representations which explicitly are about the cognitive system in question. Thus, explicit self-knowledge arises from the insight that others represent the cognitive system in the same way it represents them. Once it begins to represent its fellow cognitive systems as beings that have representations about *it*, a cognitive system will have to use an internal name for itself – thereby becoming a potential object of its own representations. The last step towards genuine self-knowledge is taken when agent-relative representations and representations containing the new internal name for the cognitive system are connected in such a way that these two kinds of representation become equivalent. For it is this move that bestows the special causal/computational role characteristic of *de se* attitudes on representations containing the system's new internal name for itself.

---

7. Cf. note 5.

References

Descartes, R. *Meditationes de prima philosophia. Œuvres des Descartes VII.* Publiées par C. Adam et P. Tannery. Nouvelle Présentation. Paris, J. Vrin, 1964-1976.

Descartes, R. *Selected Philosophical Writings*, transl. by J. Cottingham, R. Stoothoff and D. Murdoch. Cambridge: Cambridge University Press 1988.

Descartes, R. *The Philosophical Works of Descartes*, *Vol.1.*, transl. by E.S. Haldane and G.R.T. Ross. Cambridge: Cambridge University Press 1911.

Locke, J. *An Essay Concerning Human Understanding*, ed. by P.H. Nidditch, Oxforf: Clarendon Press 1975.

Lowe, E.J. 2000 *An Introduction to the Philosophy of Mind*. Cambridge: Cambridge University Press.

Perry, J. 1979 "The Problem of the Essential Indexical", in J. Perry, *The Problem of the Essential Indexical and Other Essays*, Oxford: Oxford University Press 1993, 33-52.

Perry, J. 1998 "Myself and I", in Marcelo Stamm (ed.) *Philosophie in synthetischer Absicht (Festschrift für Dieter Henrich)*, Stuttgart: Klett-Cotta, 83-103.

Rosenberg, J.F. 1986 *The Thinking Self*, Philadephia: Temple University Press.