De Se Attitudes and Semiotic Aspects of Cognition

Erich Rast*

Overview

In this article, I will re-examine some of the classical puzzles for de se attitudes that have been laid out by Hector-Neri Castañeda, David Lewis and John Perry in various articles and compare them with Jackson's Knowledge Argument. The origins of these puzzles go further back to work by Russell on egocentric particulars, by Frege in 'Der Gedanke', Wittgenstein's considerations on subject-uses of *I* in the Blue Book, and work by Roderick Chisholm. Nevertheless, it is fair to say that de se puzzles gained widespread popularity only later due to publications by Castañeda (1967), Perry (1977, 1979), and Lewis (1979).

The article starts with a survey of well-known de se puzzles: Perry's supermarket example, his Rudolf Lingens example, and David Lewis's Two Gods thought experiment. I will then discuss Jackson's Mary example, which bears a striking similarity to de se puzzles. After this primarily exegetical part, I will address the question regarding what these puzzles have been taken to show and what they really show. My central thesis is that typical de se puzzles reveal an important and epistemically irreducible aspect of thinking, but do not allow for any conclusions regarding physicalism and the Mind-Body problem. As I will argue, there is a special kind of introspective knowledge, the existence of which is fully compatible with physicalism and this special kind of

* An earlier version of this paper was presented at the conference 'Metaphysics of the Self' at the Institute for the Philosophy of Language (IFL), Lisbon, in December 2009. I would like to thank Klaus Gärtner, Jorge Gonçalves, Franck Lihoreau, Daniel Ramalho, António Marques and Dina Mendonça for fruitful suggestions and commentaries. This research was conducted under a postdoctoral fellowship from the Portuguese *Fundação para a Ciência e Tecnologia*.

knowledge results from the fact that all sorts of episodal thoughts plays a particular role in thinking and cannot be replaced by other kinds of thoughts. On the basis of this insight, I will suggest a trivializing interpretation of de se puzzles: A thought of a certain type, say α, is a necessary condition for the occurrence of a corresponding α-behaviour or action, simply because episodal thought tokens are divided into distinct classes according to the role they play in cognition. Correspondingly, it is highly unlikely that thoughts of type α, which present subjective experiences in cognition to the one currently thinking, could play the same role in that person's thinking as thoughts of another type, say β, which present physical knowledge to the one currently thinking, and vice versa. This does not mean, however, that physical knowledge cannot explain thoughts of type α or that instances of α and β belong to different ontological categories.

Before going on, some terminology must be clarified. Talk about thinking is often ambiguous between thinking in an episodal sense and dispositional thinking in the sense of having the ability to entertain certain thoughts or having a dispositional belief that *p* for some embedded proposition *p*. In what follows, I have the former in mind when talking about thinking here; this kind of thinking might also be called *cognition* in order to set it apart from the dispositional reading. Cognition is time-bound and actual. In contrast to this, in what follows belief and other propositional attitudes should be understood in the common dispositional sense. When an agent thinks (viz. cognates) that *p* this means that he currently entertains a *p*-thought or is having a *p*-thought. In contrast to this, when someone believes that *p* he has a disposition to act as if *p* were the case. Without further qualification the terms *episodal thinking* and *cognition* leave the question open whether the respective agent having the thought endorses the embedded proposition or not, i.e. whether she considers the embedded proposition true or not. I shall understand these terms in their non-philosophical sense in what follows, according to which the agent indeed takes the embedded proposition to be true. Understood in this sense, when someone thinks that *John is 32 years old* he also dispositionally believes that John is 32 years old and does not just 'think the thought' without being committed to its truth. The converse does not hold. From the fact that someone dispositionally believes that *p* it does not follow that he thinks that *p* at a certain time. I will often talk about dispositional knowledge instead of belief, since the kind of beliefs about one's own thoughts I will discuss result in some form of introspective knowledge as long as the person in question does not suffer from serious mental illness. My main points could be made entirely in terms of belief instead of knowledge, though, and so not too much weight should be given to this terminology.

Puzzles of De Se Attitudes

The supermarket example laid out by Perry (1979) is one of the clearest cases for the de se attitudes:

> 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. But I seemed unable to catch up. Finally it dawned on me. I was the shopper I was trying to catch. (Perry, 1979, p. 3)

This scenario has two different, though related aspects. On one hand, Perry might cognate in various ways about himself without realizing that his episodal thoughts are about himself. For example, he might think that someone/the only bearded philosopher in a Safeway store west of Mississippi/John Perry/this man (+pointing gesture) in the mirror is making a mess without realizing that it is he himself whose sack of sugar is torn. As long as John Perry is sufficiently amnesiac, doesn't remember his name or that he hasn't shaved himself in the morning, and doesn't recognize himself in a mirror, none of these ways of thinking about himself seems to explain his behaviour, until he starts to think: It is *me*, who is making a mess, *I* am producing a trail of sugar! On the other hand the example also has a linguistic aspect. The different ways in which John Perry might linguistically realize his episodal thoughts don't seem to have the same explanatory power for his action than Perry's utterance of *I am making a mess.* For example, in order to express the same thought using his proper name, John Perry would have to additionally believe that he is called *John Perry*, which he doesn't believe

in the given scenario. This has led Perry and others to the conclusion that there is an essential reading of the first-person indexical that is irreducible with respect to its power for explaining certain changes in behaviour and actions.

Another well-known example can be found in Perry (1977):

> An amnesiac, Rudolf Lingens, is lost in the Stanford library. He reads a number of things in the library, including a biography of himself, and a detailed account of the library in which he is lost. He believes any Fregean thought you think might help him. He still won't know who he is, and where he is, no matter how much knowledge he piles up, until that moment when he is ready to say, 'This place is aisle five, floor six, of Main Library, Stanford. I am Rudolf Lingens.' (Perry, 1977, p. 492).

Again, the problem in this example is that it always seems to be conceivable that the epistemic agent in question gathers all kind of knowledge about himself from external evidence, but only when he realizes that this information is about *himself* will he be able to act accordingly. So it appears as if, from an epistemic point of view, the agent learns something when he realizes that he is John Perry or Rudolf Lingens respectively, he himself is making a mess on the floor or he himself is located at aisle five, floor six, of Main Library, Stanford, and so on. Lewis (1979) has provided a famous variant of these examples that, albeit being highly contrived, is rather instructive from a logical point of view, because it is spelled out in terms of possible worlds semantics for propositional attitudes like dispositional belief:

> Consider the case of two gods. They inhabit a certain possible world, and they know exactly which world it is. Therefore they know every proposition that is true at their world. Insofar as knowledge is a propositional attitude, they are omniscient. Still I can imagine them to suffer ignorance: Neither one knows which of the two he is. They are not exactly alike. One lives on the top of the tallest mountain and throws down manna; the other lives on top of the coldest mountain and throws down thunderbolts. (Lewis, 1979, pp. 520–521)

The idea behind this thought experiment is as follows. In epistemic logic based on normal modal logic an agent's epistemic state is represented by the set of possible worlds, i.e. maximally truth-making doxastic alternatives, that are accessible from the actual world by an accessibility relation for that agent and the respective kind of propositional attitude.

Suppose the belief set Γ is the set of possible worlds accessible by an agent's current belief relation from the actual world. Consider now a case when the agent is in doubt whether p is the case or not. In a possible worlds setting, this boils down to saying that there are some w_1, w_2 in Γ such that p holds at w_1 and $\neg p$ at w_2. When the agent learns that p is the case, the revised accessibility relation will yield a set of possible worlds from which all $\neg p$ worlds have been removed. The more an agent learns about the universe, the smaller becomes his belief set. If only one world is left and as in Lewis's scenario that world is the actual world, then the agent is omniscient (see Figure 1 for illustration). Whatever he can learn about the state of nature he has already learned. Nevertheless, Lewis claims that for example Zeus in this scenario still doesn't know that he himself is Zeus, that he himself is sitting on the tallest mountain and is throwing down manna, and so on.

Lewis has devised this scenario to illustrate a limitation of the usual truth-conditional semantics for dispositional attitudes based on modal logics with Kripke or Scott-Montague semantics. He concluded from the thought experiment that possible worlds alone do not suffice to represent what an agent might learn, hence do not suffice to represent epistemic states and corresponding dispositional attitudes in general, and for this reason he proposed a more fine-grained model in which having a relational attitude is modelled as the self-ascription of a property by the respective agent. Since properties in Lewis's view are more fine-grained than sets of possible worlds, more attitudes can be distinguished by the property-ascription view and so his approach can adequately represent the assumption that the two gods in the scenario might have different epistemic attitudes despite being omniscient about all the 'external' facts. Lewis's solution was not very appealing to many philosophers, and rightly so, because he did not present enough details about the properties he had in mind. Lacking a detailed and positive metaphysical account of properties his suggestion remains unsatisfactory. Fortunately, many other solutions to the Two Gods puzzle have been proposed during the past few decades, and, technically speaking, any way to make dispositional belief more fine-grained will do as long as one merely strives for descriptive adequacy with respect to de se puzzles. One of the simplest solutions, which was also discussed by Lewis (1979), is using centred possible worlds. Instead of bare possible worlds ordered pairs

consisting of an agent and a possible world are taken as semantic base entities. With this small change the lack of omniscience of the two gods can be easily represented. For example, Zeus's belief set $\{\langle w, Zeus\rangle, \langle w, Jahwe\rangle\}$ would contain two tuples, one containing himself and the other one containing Jahwe as long as he hasn't realized that he himself is Zeus. When he learns that he himself is Zeus he removes the pair containing Jahwe from his belief set (see Figure 2). There are many other approaches to hyperfine-grained attitudes that may be used for dealing semantically with Lewis's scenario: structured propositions from von Stechow (1982) and Cresswell (1985), approaches based on substituting reified propositions for truth-values as in Thomason (1980), using impossible worlds as in Hintikka (1975), and property theory of Bealer & Mönnich (1989). To make a long story short, it is fair to say that de se puzzles nowadays no longer pose any particular technical challenge to a logician who is interested in a descriptively adequate truth-conditional semantics for belief and similar attitudes.

However, none of these approaches give a satisfactory answer to the question of what exactly the agent learns when he realizes that he himself is called by a certain name or satisfies a certain property. The power of Lewis's example can be illustrated further by thinking of possible worlds for a moment in an old-fashioned way as possible constellations of matter and the forces that hold it together. By assumption, in the example both gods know everything there is to say about the constellation of matter in the universe and the forces that hold between matter. Still, so it is claimed, they can learn something. Consequently, whatever there is left for them to learn cannot be an aspect of matter and the forces between that matter. As Stalnaker (2004) remarks, there is a striking similarity between de se puzzles of this kind and the Knowledge Argument by Jackson (1982, 1986), which was originally intended to show that qualia are not physical and therefore physicalism is false. Jackson (1986) has laid out this example in the following, much cited passage:

> Mary is confined to a black-and-white room, is educated through black-and-white books and through lectures relayed on black-and-white television. In this way she learns everything there is to know about the physical nature of the world. She knows all the physical facts about us and our environment, in a wide sense of 'physical' which includes everything in *completed* physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon this, including of course functional roles. If physicalism is true, she knows all there is to know. [...] It seems, however, that Mary does not know all there is to know. For when she is let out of the black-and-white room or given a color television, she will learn what it is like to see something red, say. (Jackson, 1986, p. 291)

Just like the gods in Lewis's scenario Mary is omniscient about the external world, yet she doesn't know what it feels like to experience red, before she has been exposed to a red object. In de se puzzles the agent still needs to learn that it is he himself/she herself who has a certain property or is called by a certain name. According to the Knowledge Argument, the agent is also supposed to learn something in addition to her physical knowledge about the world, namely the way in which a certain experience feels to her. So do these thought experiments indicate that there is something nonphysical like an irreducible *I*-thought or a particular red quale? As I will argue in the following section, the answer to both questions is No. Neither de se puzzles nor corresponding arguments for qualia show that the physical world is not closed or that certain phenomena in episodal thinking are not reducible to physical states. Both kinds of puzzles show, however, that there is a special sort of introspective knowledge which is fully compatible with physicalism and based on remembering past experiences or thoughts. In order to have this kind of knowledge, an agent must store an experience or thought in a way such that the retrieval of this memory fulfils a certain role in his thinking; when something else is stored instead, it might not fulfil the same role when it is retrieved. Investigating this kind of knowledge in combination with some general semiotic considerations will pave the way for a trivializing interpretation of de se puzzles.

Assessment of the Thought Experiments

It is worth noting that the Two Gods thought experiment does not strictly speaking present a positive argument for the thesis that there *are* de se attitudes. Instead of acknowledging that each of the gods lacks some

knowledge one could also take the example as illustrating that the gods do not lack any knowledge, since there are no epistemic alternatives left for them to consider. By assumption Zeus knows everything there is to know about the physical world, hence one might argue that he also knows who he is, what his name is and which properties he has, that it is himself who is currently thinking, whether he is throwing down manna or thunderbolts, and so on. *If* the mental supervenes on the physical, the defender of physicalism might argue, *then* Zeus also knows everything there is to know about his mental life and consequently does not lack any knowledge about his own episodal thoughts either. From this point of view Lewis's premise that the two gods lack knowledge could be regarded as false. Let us call this the *denial reply.* It does not seem to be very convincing. On the one hand, the thought experiments have an intuitive power that lures us into believing that the agent can indeed learn more. Using thought experiments in this way has rightly been criticized by Dennett (1991) as intuition pumping in an area where our intuitions may be utterly misleading. On the other hand, the denial reply neglects an important distinction between the static theoretical knowledge an agent has at a certain time and the knowledge an agent might gain when he or she has an insight about him/herself. Nothing in the description of the scenario warrants that one of the gods actually has the insight that it is he himself/she herself who has such-and-such properties.

Jackson's Mary example has been criticized by Dennett (1991) in a similar fashion as in the denial reply to the Two Gods example. According to Dennett, Mary does not learn anything new when she is exposed to a red object for the first time and we only find this claim counter-intuitive because we cannot adequately imagine what it means to have exhaustive knowledge about the physical world. Many other replies have been given to the original argument, ranging from a critique of 'intuitions' to the denial of one or more of its premises, and Jackson himself has by now converted to physicalism. For example, according to the well-known ability hypothesis Mary does not gain new knowledge but an ability – see Lewis (1983, 1988/1990), Nemirow (1980, 1990), and Churchland (1985, 1989), cf. Coleman (2009) for a comprehensive critique. Others such as Conee (1994) and Tye (2009) have defended the so-called acquaintance hypothesis according to which acquaintance with certain objects or experiences cannot be explained in terms

of knowing that, or abilities. I will in the following paragraphs lay out a position according to which (a) in the classical de se puzzles and in Jackson's Mary example the agent in question learns something new in the sense of gaining new knowledge, but (b) neither de se scenarios nor Jackson's Mary example speak against physicalism, (c) the examples are described incorrectly, since the world changes in each of them, and (d) episodal thoughts of a certain kind can trivially not be substituted by episodal thoughts of another kind and dispositional knowledge cannot replace episodal thinking. In a nutshell, what I claim is that in order to know the episodal thought that you've been thinking (or are just thinking), you need to think it first, and thinking a certain episodal thought is trivially different from an explanation of that thought or an episodal thought that explains that thought.

To back up these claims let me start with an analogy. Imagine a little robot toy that contains sensors that react to light. A sensor function $s: T \rightarrow \{l, r, f\}$ from time to three values works as follows: When there is more light to the left than to the right, the sensors return l; otherwise if there is more light to the right than to the left they return r; otherwise they return f. The robot automatically stores each sensor input in its memory and can also perform three actions: move forward F, move forward left L, or move forward right R at a time t. The device is programmed as follows: (1) Store $s(t)$ in memory. (2) If $s(t)=r$ then R, else if $s(t)=l$ then L; otherwise F at discrete times t measured by an internal clock. If designed correctly, this little device will follow the strongest light source until it bumps against something, malfunctions, or its battery runs out. In its memory a certain sequence is stored, and which one depends on its environment. If, for example, the robot is never exposed to light from its left, it may go in circles to the right and the symbol l will never be stored in its memory. Now suppose your colleague from the local A.I. department wanders by and laments that this device is much too simple. He upgrades the robot's software and hardware to allow for running complex programs on it, including conditional actions based on the contents of the memory and devising simple plans such as 'If the last sequence of actions was *llrll*, then R.' Despite this gain in abilities and 'cognitive powers', as long as the robot is never exposed to light from the left none of its routines that introspect l states and act conditionally upon them will do anything. Suppose further that

your colleague from the A.I. department becomes famous by inventing a rather complicated program that, when hooked up with sensors and running on suitable hardware, becomes conscious and self-aware. (As a philosopher, you remain rather sceptical about this conjecture but the dispute ends in a stalemate: neither can you falsify his claim that the running program is conscious nor can your colleague falsify your claim that it is not conscious.) Out of generosity your colleague upgrades your robot toy, uploads the program, and immediately once the program is started the robot begins to explore the world. Being equipped with sophisticated A.I., this robot bears some resemblance to RoboMary described by Dennett (2005, pp. 122–129), but it illustrates a different point here. The point is this: Even though our robot is now conscious, self-aware, and capable of sophisticated planning and other higher cognitive abilities, what it can experience, learn, and do is restricted by its environment. If you mount a strong light on its right side, the robot will never store the symbol *l*. Since the robot is conscious, this means that it will never have an *l* experience, or, to put it in other words, the robot will have no possibility of knowing how it feels like to have an *l* experience.

This example shows two things. First, it is obvious that the robot does not lack any ability when it is never exposed to light from the left. It has the same abilities when it is never exposed to light from the left as it has when it is exposed to light from both sides and does not lack any know-how in either case. So the so-called ability reply to the Knowledge Argument does not seem to be convincing if the above example illustrates basically the same problem as Jackson's Mary example. Secondly, there does seem to be a sense in which the robot learns something new when it is exposed to light from the left for the first time, even though its blueprint and the software running on it doesn't change. It has a certain experience that it has never had before and can now store the fact that it has had this experience in its memory. It acquires memory of a past experience, which is, by definition, a learning process.

Keeping the robot in mind, let us return to the Mary example. Does Mary learn something new when she is exposed to a red object? The answer is clearly Yes. She can now remember an experience that she didn't have before. Does that mean that qualia are not physical or that there is some irreducible phenomenal knowledge? The answers which one gives

to these questions depends on the stance one takes about red experiences viz. *l*-experiences. Someone who thinks that the robot can only have an *l*-experience if it is conscious, i.e. consciousness is a necessary condition of having an *l*-experience, would without doubt be inclined to say the same about Mary. In this view, which has been criticized by Dennett and others for its reliance on unstable intuitions, Mary learns something new when she is exposed to red, but only if she is conscious. In this point of view, machines cannot be conscious as they merely follow a fixed set of instructions. So when our robot acquires knowledge and eventually becomes just as brilliant a scientist as Mary, it will still not have an *i*-experience even if it is exposed to light from the left. Notice, however, that neither of the thought experiments supports this view or its opposite. Certain types of computational structures, when running on some hardware, or certain types of physical structures could be conscious without anyone ever knowing for sure that they are. Suppose this were the case. Then there would *still* be a difference between Mary's epistemic state before she has been exposed to red and afterwards and the robot's epistemic state before it has been exposed to light from the left and afterwards.

Moreover, suppose the robot was omniscient about the physical world and *not* conscious. Still, the robot's epistemic state would be different before and after it has been exposed to light to the left, because it can store *l*-input from the sensors, which is a learning process, and being able to remember a past input is a kind of knowledge. But in order to be able to remember a past input or experience the respective agent needs to have had it. So there is indeed a special kind of knowledge at play that one might label as 'phenomenal', but the conditions for acquiring this knowledge are entirely independent from the question whether qualia are physical or not and from the question whether an agent is conscious or not. *If* qualia are nonphysical and exist, then the occurrence of a nonphysical phenomenon is a necessary condition for having knowledge of qualia experiences, but the Knowledge Argument does not show this. The Knowledge Argument merely shows that certain knowledge can only be acquired by having certain experiences, be these reducible to physical structures or not.

A very similar point can be made about the Two Gods example. Since each of the gods is omniscient about the physical universe, they do know who they are and are *able* to identify themselves. However, there is

still a sense in which each of them lacks knowledge, as long as they do not make use of their ability. For instance, unless Zeus thinks the episodal thought *I am Zeus* he cannot get to know that he has been thinking that thought. Regarding his dispositional belief this means that the occurrence of the episodal thought *I am Zeus* is a necessary condition for his having the dispositional de se belief that he himself is Zeus. The thought has to occur in actual cognition first, and only after it has occurred can the agent have the corresponding dispositional de se attitude. If this is so, then there are two ways to interpret the puzzle. First, one may claim that as a consequence of having complete theoretical knowledge about the universe an agent will invariably come to episodically think corresponding thoughts of the form *I am F.* Since we are talking about arbitrary properties *F* that the agent possesses, this view does not seem to be very compelling. If at all, it only makes sense under very strong rationality assumptions and when the agents in question are highly idealized like the two gods. Since no actual, resource-bound agent can have infinitely many episodal thoughts, and thus every actual agent lacks knowledge about herself, another response seems to be more appropriate. According to this view, the complete theoretical knowledge of the agent in question does in itself not warrant that she has corresponding thoughts of the form *I am F*. *If* the agent had such a thought, *then* she would be able to recognize that she has had it even from a purely 3rd-person perspective, provided that episodal thoughts are entirely reducible to certain kinds of physical structures, but since having the dispositional knowledge does not cause or otherwise warrant that the agent actually has an episodal thought of the kind needed for having a corresponding de se belief, she might not recognize that she herself is *F*. If, on the other hand, episodal thoughts are not entirely reducible to certain kinds of physical structures, then an agent that is omniscient about the physical world might not recognize from a 3rd-person perspective that she has had that thought. In both cases, however, the agent can only recognize that she has had such a thought after she has actually had the thought. Again, the puzzle does not decide anything about the ontological status of the mental. All it says is that you need to actually *think* certain thoughts in order to be able to retrieve from memory that you have thought them, and, as in case of the robot, the ability to retrieve memories of past mental events, be they ultimately physical

or not, can be considered a form of knowledge. This sort of knowledge may be called *introspective knowledge.*

Semiotic Aspects of Cognition

From what has been said so far, a number of conclusions can be drawn. Firstly, the Two Gods and Jackson's Mary example are fully compatible with physicalism and neutral about the question whether physicalism holds or not. Secondly, the descriptions of the examples are incomplete. Suppose physicalism is true and Mary is experiencing something red. Then her brain state changes, too, and consequently her previous knowledge about the physical world is outdated. Likewise, if physicalism is true and an agent has the episodal thought *I am F,* then having this episodal thought corresponds to a change in the agent's brain state, rendering his previous knowledge incomplete. A static picture of the universe, as is presumed by using some simple, tenseless possible worlds semantics for propositional attitudes, is not entirely adequate for describing these scenarios. When for example Mary's knowledge is updated according to the changes in the world after she has been exposed to a red object for the first time, she would be able to deduce from her physical knowledge alone that she has had a red experience. Nevertheless, as simple as this may sound, the fact that she has had a red experience is a necessary condition for her to realize, on the basis of physical knowledge about the world only, that she has had a red experience. As I will lay out in more detail below, even under the premise that physicalism is true, recognizing the red experience by means of measurement and physical knowledge and subsequently forming knowledge about its recognition by remembering it is not the same as remembering the red experience itself.

This position is similar to what has been proposed by defenders of what Nida-Rümelin (2009) calls the new knowledge/old fact view but with one rather crucial difference: in the present view, the fact *is* new, although it will be possible to fully explain it in physical terms if physicalism is true. If on the other hand physicalism happens to be false,

then having a red experience trivially amounts to a new phenomenal fact, namely the fact that the person in question had *this* particular red experience. In both cases, the world changes and so the knowledge of the omniscient agent in question needs to be updated. What about the updated knowledge then? Does the updated dispositional knowledge suffice to explain a particular red experience or an agent's insight that he himself or she herself has a certain property? The correct answer is: Why not? Why should the updated knowledge not suffice to *explain* the respective experiences provided that physicalism is true? Explaining in this sense means nothing more than knowing the conditions under which the agent in question has the respective experiences and therefore is able to form introspective knowledge about them. However, there is no reason whatsoever to believe that dispositional knowledge of an agent or someone else about an agent's experiences or thoughts can substitute or replace in any way the agent's having certain episodal thoughts that present experiences, insights or memories thereof to him/herself in his/her cognition. Even under the assumption that certain episodal thoughts directly correspond to dispositional, physical knowledge about a particular red experience *this red*, i.e. when we would confine further considerations to whatever corresponds to dispositional knowledge about *this red* in actual cognition – like the episodal thought that the episodal thought *this red* is such-and-such –, there is no reason whatsoever to believe that these kinds of episodal thoughts could or should play the same role in cognition as the thought *this red* that an agent has when he is confronted with a red object or the thoughts he might have when he imagines or remembers a red object. Likewise, various ways of identifying oneself in actual cognition in 3rd-person ways might not play the same role in cognition as the thought circumscribed as being of the form *I am F* that occurs in cognition when an agent realizes that he has the property *F*.

The resulting perspective on de se attitudes is trivializing. Certain episodal thoughts are necessary conditions for a certain kind of behaviour, whereas other sorts of episodal thoughts are not – no matter what we know about episodal thinking and how it should be described. An explanation of our robot's inner workings does not substitute the machine's actual having and processing certain sensor inputs or fetching data from its memory even if the robot itself processes this explanation

and accurately measures its own state. Correspondingly, Mary's knowledge about the physical world cannot substitute her having a red experience. Generally speaking, episodal thoughts of a certain sort play a certain role in thinking that thoughts of another sort might not. This thesis should be relatively uncontroversial, yet it has been overlooked consistently in the assessments of respective thought experiments.

Episodal thinking has another important property that is relevant for the assessment of the examples: it does not work like a language. Too see why, consider the robot example again. In a well-functioning robot an *l* symbol occurs when the left sensor receives more input than the right sensor, and correspondingly for the processing of *r* and *f* symbols. The symbols are stored in a memory module of a given capacity. Suppose now that the robot's program says 'When the memory contains the sequence *lll*, remove it from memory and execute R for the next two time units.' Another part of the program could possibly use another symbol, say *a*, to refer to a sequence *lll*. However, in order for the symbol *a* to have the same effect as in the above programming instructions it has to be assigned in an appropriate way to sequences of *lll*'s stored in memory; no matter how this connection looks like in detail the robot must remove three *l*'s from memory and give the *R* signal for the motors to go right for the next two time units in order to run the same program on the basis of the alternative representation *a*. In this respect, although the choice of the symbol *l* is completely arbitrary, a sequence of three *l*'s in memory is the most basic representation of three inputs from the left sensor on this machine with respect to the particular program running on it. While it would be possible in this example to replace three *l* symbols with a single one such as *a*, which is stored whenever the left sensor has received three subsequent inputs in a row, this strategy only works because the program only acts upon three received *l* symbols. It does not work in general. Suppose the robot gets the following, more complicated instructions: 'Whenever there is a sequence of *n* symbols *l* followed by a sequence of *n/2* symbols *r*, remove the whole sequence and drive right for *n* time units.' This rule involves the counting of sensor inputs, and for this purpose at some level of description the robot must store the number of subsequent *l* symbols in memory. To store the number of subsequent *l* inputs, one may for example store *n* symbols *l* for *n* inputs from the left sensor. Alternatively, binary states could be

used. The number *3* is then represented by setting two states to *on* (see Figure 3). No matter which representation for *n* inputs of a kind is chosen, though, in order to run the program successfully it must be able to produce motor output for *n* time units. Taking into account the complete device, including the software running on it, we may thus say that the desired representation of *n* sensor inputs *encodes* the number *n*.

Most symbols do not encode what they stand for. Consider for example the Arabic digit *3*. In order to be able to represent three things this symbol must be mapped to three signals, objects, or other entities of a suitable kind. The symbol *3* does not itself exemplify what it stands for and it may thus be called a *purely representational symbol*, or, in the terminology of Langer (1951), a *discursive symbol*. From these symbols Langer distinguishes *presentational symbols* such as the three dots in Figure 3. Presentational symbols exemplify what they stand for in addition to representing it. In other words, they *encode* directly at least certain aspects of what they represent. Notice that the two binary gates are a tricky case. They seem to be more presentational than the digit 3, yet without a rule that maps the *on*-state of one of them to two objects or signals and the *on*-state of the other one to one signal or object, the gates cannot for themselves present three objects directly. If the mapping mechanism is included in the description of the symbol, then it may be considered presentational but otherwise it is purely representational just like the symbol *3*.

I do not claim that this distinction can be generalized easily to cases that are more complicated than the encoding of natural numbers. A simple isomorphism will most likely not suffice as an explanation of more complex presentational symbols and iconicism is one of the most vexing problems of semiotics. Instead of grounding the desired distinction on the notion of an isomorphism or making similar attempts to give a precise account that would require much more support, I will only assume some less stringent and more tentative definitions in what follows. Let us speak of a purely representational sign if the connection between sign and the signified is conventional and (in this sense) arbitrary. Let us, following H.N. Castañeda to some extent, assume in contrast to this that a presentation encodes finitely many aspects of what it represents in a way that makes the connection between the sign and the signified not just conventional and arbitrary. (The reason for allowing such a vague

definition here is that an explanation of the exact meaning of the phrase 'not just conventional and arbitrary' should be left to neurophysiology if physicalism is true and psychology if physicalism is false. That at least some such presentations exist has been illustrated by the example of finite presentations of natural numbers.) Given this distinction, it is clear that languages are purely representational, for it is well-known that even allegedly iconic expressions like onomatopoeia in natural languages are conventionalized. Cognition, on the other hand, cannot be purely representational. If cognition worked like a language and were also purely representational, then each thought token would stand for something else. As I briefly laid out in Rast (2007, pp. 270–275), this view would lead to an infinite regress, since an agent would then have to check to which entity a particular thought token refers in order to 'understand' the token, and so on for the symbols the token refers to, for the symbols the referent of the token refers to, etc., until a presentational symbol is encountered. This modern Homunculus problem shows that episodal thinking cannot in general be only representational; it must at some point involve presentations of objects, experiences, and so on. Moreover, as I've mentioned before and as H.N. Castañeda has emphasized throughout his work (see for example Castañeda, 1989, 1990), humans are finite and therefore presentations of objects or experiences in episodal thinking must be finite, too. Consequently, when an agent thinks a thought that may be circumscribed in public language as *I am F* for some property *F*, the *I*-presentation of herself in her cognition encodes finitely many aspects of herself and she attributes the property *F* to this presentation. Likewise, a presentation of a particular red experience in cognition presents finitely many aspects of a particular red object or a region of an agent's visual field in her cognition. If physicalism is true, these presentations are encoded and processed directly by an agent's brain – an open, analogue, massively parallel, and presumably non-deterministic computational system – and the result of such a computation is itself finite. But even if physicalism is not true, there does not seem to be any alternative to the popular view that episodal thinking is finite.

Although not the central tenet of this article and somehow independent of the previous considerations, I would briefly like to discuss the popular yet sometimes misunderstood thesis that episodal thinking is also computational in nature, which further closes the gap between

the robot and the Mary example. By 'computational' I mean that the mind or brain can within the limits of its finite resources compute only functions that are representable by terms of the λ- or π-calculus and their reduction to canonical form, or another suitable sequential or parallel Turing complete formalism. While there are alternatives to this view, they are hard to reconcile with what we know about the physical world. First of all it must be noticed that the brain or mind can compute certain computable functions within its physical limits. It is therefore not a good idea to claim that the brain or mind is not at all a computational device. So to reject the computational model one has to claim that the mind/brain is higher than a Turing machine in the hierarchy of computational systems. Hypothetical devices that can solve problems that a Turing machine cannot solve are called hypercomputers, and there is a whole hierarchy of them. A relatively weak hypercomputer would for example be the 'accelerated Turing machine', which is based on ideas already considered by Russell. An accelerated Turing machine completes its first step during computation (such as moving the head on the tape, printing a symbol to tape, etc.) at time 1, the second step at 1+1/2 time units, the third one at time 1+1/2+1/4, and so on. When it reaches the limit at time 2 it might have solved a task that provably no ordinary Turing machine can solve. There are more credible descriptions of hypercomputers, but none of them seems to be fully compatible with current physical knowledge (see Lokhorst, 2000). Other alternatives such as Penrose (1989, 1997) are more elaborate, but don't offer any convincing explanation of how the mind might work. Yet other alternatives such as claiming that the mind is neither computational nor hypercomputational nor based on quantum-physical phenomena amount to sheer mysticism. Now the fact that the computational model seems to be the best explanation of how episodal thinking works does not mean that it is the right explanation. The issue is clearly undecided. Still, the best explanation is the best explanation unless someone comes with a better one, and we should prefer the computational model unless someone comes up with direct counter-evidence to it. Notice that even if physicalism were false a computational model of the mind would still be more attractive than any of the alternatives mentioned; in that case, however, some mysticism could not be avoided due to the problem of explaining mind-matter interaction.

If the computational model is basically the right picture of how episodal thinking works, then invariably certain episodal thoughts fulfil a different role than others within the computational system as a whole. Although some episodal thoughts may in principle fulfil the same role as others if they are ultimately linked to the same sensor inputs and motor outputs, as the *a* versus *lll* example has illustrated, it is not very likely that a particular *this red* presentation or a particular *I am F* thought in human cognition could be replaced by some radically different thoughts, such as those that represent some realization of dispositional knowledge in cognition, *and still fulfil the same role within the given computational system.*

What conclusions can be drawn from these general considerations concerning de se puzzles? Let us return to Perry's supermarket example mentioned at the beginning of this article. In the thought experiment certain kinds of thoughts are connected to a certain kind of behaviour. For example, John Perry looks at himself in the mirror without recognizing himself, thinks what may be circumscribed in public language as *This guy is making a mess* and attempts to follow the person in order to tell him that his sugar package is damaged. Suddenly he thinks *I am making a mess,* which must here be understood as a rough and deficient natural language circumscription of what actually goes on in his brain (or mind, if physicalism is false) when he has the insight and correspondingly starts to clean up the sugar. According to the view I have suggested, this is so because a particular finite presentation in John Perry's cognition lead to a particular behaviour, whereas another presentation leads to another kind of behaviour. The connection of this phenomenon to natural language is loose, though. Take any double-vision puzzle like Quine's Ortcutt example in Quine (1956) Perry's ship example in Perry (1979, p. 483), or Richard's phone booth example in Richard (1983). Being forced to base their judgments on finite presentations of objects, people use certain expressions for some presentations and others for other presentations without realizing that these presentations present the same object in different ways to them. Hyperfine-grained belief can be used to model such cases, but outside of attitude ascriptions the respective natural language expressions retain their public language reference and refer to objects with infinitely many properties. As countless discussions about propositional attitudes have shown, even when the

respective natural language expressions occur inside attitude ascriptions it is controversial whether hyperfine-grained attitudes should be used to truth-conditionally encode an agent's tendency to use certain expressions as opposed to others for certain presentations in cognition. Language is purely representational and public, whereas cognition is predominantly presentational, private, and its intricacies are mostly unknown at the time of this writing. For that reason, mappings between these symbolic systems remain inaccurate and leave room for variations due to different theoretical goals and modelling purposes.

Things look different, however, when one is interested in logical representations of epistemic states of thinking or computing agents independently of natural language semantics. Certain thought tokens form equivalence classes on the basis of the role they play within the computational system as a whole, including the program currently running on it, and tokens from one equivalence class can by definition not be substituted by tokens from another one. Thus, when instances of certain classes of tokens are stored for later retrieval they play an essential role in the constitution of introspective knowledge of a certain sort that can trivially not be replaced by knowledge based on instances of another class of thought tokens within the computational system. De se attitudes and the special status of subjective experiences within an agent's thinking as opposed to 3rd-person knowledge about them are symptoms of this general property of computational systems. So from an epistemic point of view hyperfine-grained attitudes seem to be unavoidable. However, being closely tied to purely representational formal languages, sometimes even to their syntax, existing accounts of hyperfine-grained attitudes lack a certain explanatory adequacy concerning these phenomena even when they are descriptively adequate. An explanatory satisfying account of hyperfine-grained attitudes for humans would have to be based on a comprehensive theory of presentational, episodal thinking in general, and no such theory is available at the time of the writing of this paper. If what has been said above is correct, according to such a theory episodal thinking is radically different from what we are used to calling a language.

Finally, some things have to be said about the talk about 'necessary conditions' in the previous sections. I have *not* laid out any principal reasons why thought tokens that represent physical knowledge about

the world could not substitute thought tokens for subjective experiences or tokens that present thoughts like *I am F* for a property *F* in the sense that they might, under certain circumstances, fulfil the same role in episodal thinking as the thoughts they are considered to replace. There is no principal reason why within a computational system one sort of signals could not play radically different roles dependent on the overall state of the system, and I have only claimed that assuming such a dual role for certain thoughts in human thinking is an implausible attempt to explain away introspective knowledge. My reply to Dennett and the denial reply to the Two Gods puzzle is not that it is infeasible that certain theoretical knowledge could necessarily not play the desired role in thinking; my reply is rather that it actually doesn't. How then, as one might ask, does this support the much stronger thesis that certain thoughts are a necessary condition for certain actions? The answer is that this depends on the way thoughts are categorized. The way typical de se puzzles are set up no thought other than *I am F* plays the same role for subsequent *I*-behaviour, even though another thought such as *John Perry is F* could in principle result in the same behaviour. *I*-thoughts lead to *I*-behaviour by definition, and, as one might continue, *John Perry*-thoughts lead to *John Perry* behaviour, *this guy*-thoughts lead to *this guy*-behaviour, and so on. If someone would claim that in the supermarket example a *John Perry*-thought might lead to the *I*-behaviour, a defendant of de se attitudes could reply that this is not possible, because thoughts that result in the *I*-behaviour, i.e. in the agent having the insight about themselves respectively described by the scenarios, are *I*-thoughts. They fulfil a role in episodal thinking and memory thereof which manifests itself as the *I*-behaviour. Understood in this way, having an *I*-thought is a necessary condition for the occurrence of an *I*-behaviour, because thoughts have been classified in this way. This is the trivial aspect of the suggested interpretation of de se puzzles, but as I have shown above they also exemplify the non-trivial formation of introspective knowledge. Using the same trivial 'cognitive a priori' definition of thoughts in the Mary example would render it rather incomprehensible, because in this way of talking we would have to say that if Dennett was right then thoughts that represent an omniscient scientist's knowledge about colour vision would just *be* thoughts that represent a certain colour experience. Not even Dennett talks this way. We usually talk in another way about

cognition, namely in the way that thoughts are not individuated by their role; so in this way of talking two different thoughts might (at least in theory) play the same role, yet remain different from each other.

Summary and Conclusions

Regarding the Mary example, I have outlined an approach that Dennett (2005) calls 'thick materialism'. Mary learns something new and this new, introspective knowledge is compatible with physicalism. The particular insight of an agent about himself that is described in typical de se examples and an agent's memory of this insight likewise result in introspective knowledge. In both cases the agent stores a certain thought token in memory and by having had the particular thought and being able to remember having had it in some way the agent acquires introspective knowledge. Within a computational system certain signals or thought tokens play a certain role in the system as a whole that other thoughts might not play. If *t* is the symbolic presentation of a certain sensory input within a given computational system, then a presentation *t'* of knowledge about *t* can only play the same role within the computational system as *t* if it is suitably linked to other parts of the system in the same way as *t*. For example, when *t* consists of three symbols or signals *lll* that represent three inputs from a certain sensor and within the system at that time there are also some actual presentations *t'* of the system's dispositional knowledge about *lll* symbols, their storing and retrieval, and the various roles they can play in relevant subroutines of the computational system then under usual circumstances *t'* cannot substitute *t* in the sense that *t'* cannot play the same role as *t* within the computational system as a whole for the roles that are fine-grained enough to be of interest for a role-based explanation of the system. Concerning human thinking it seems even less likely that thoughts presenting sensory experiences could be replaced by thoughts that encode dispositional knowledge while retaining the same role within the system as a whole.

If phenomenal concepts are taken as a condition for the ability of an agent to form introspective knowledge of the kind laid out in this

article, then it follows from what has been said above that these can be understood in purely physicalist terms without resorting to ontologically or metaphysically irreducible phenomenal properties. However, there does not seem to be any need for any special conditions of this sort, since the fact that humans have thoughts of different kinds in combination with the fact that they have the ability to remember them (to some extent) suffices as a general explanation as to why humans have the ability to form introspective knowledge.

Figures

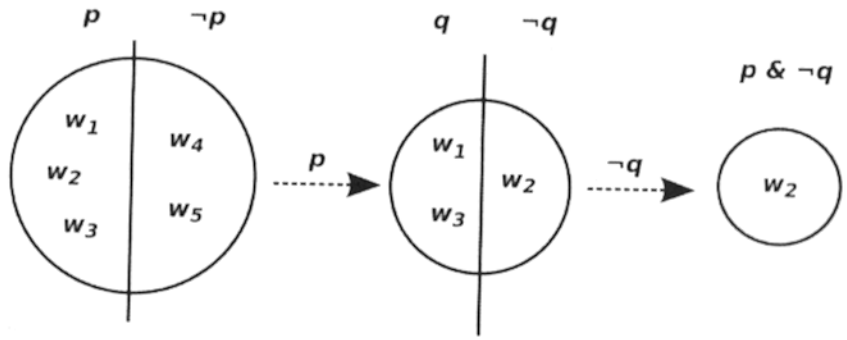

Figure 1: An agent's belief set is shrinking when he accepts new propositions.

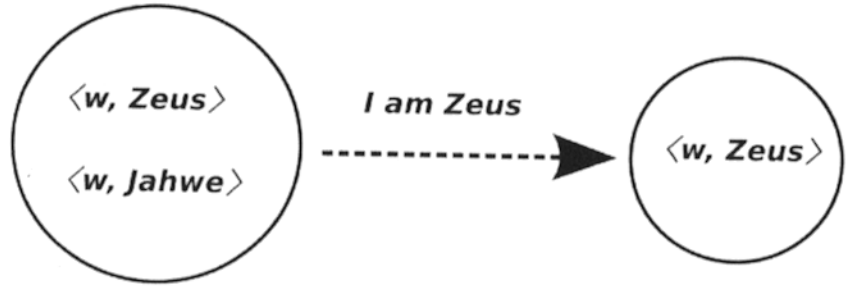

Figure 2: Using centered propositions for representing de se belief.

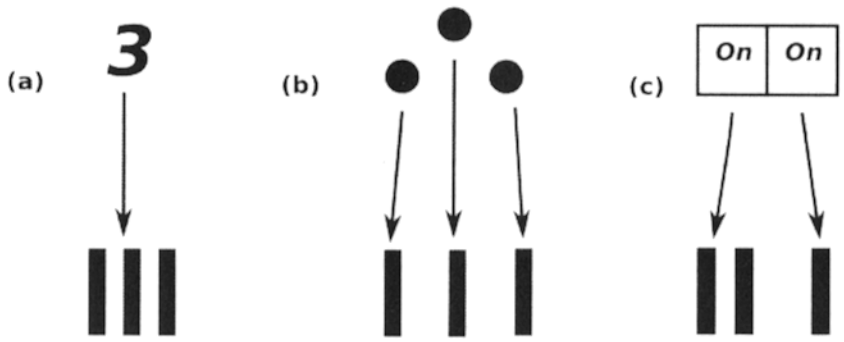

Figure 3: Different ways of representing the number 3 and their respective decoding: (a) Arabic digit, (b) three dots, (c) two active binary states.

References

Bealer, G. & Mönnich, U. (1989). Property Theory. In Dov Gabbay (ed.), *Handbook of Philosophical Logic*. Dordrecht: Kluwer, 133–251.

Castañeda, H.-N. (1967). Indicators and Quasi-Indicators. *American Philosophical Quarterly, 4*, 85–100.

Castañeda, H.-N. (1989). *Thinking, Language, Experience.* Minneapolis, Minn.: University of Minnesota Press.

Castañeda, H.-N. (1990). Indexicality: The Transparent Subjective Mechanism for Encountering A World. *Noûs*, *24*(5), 735–749.

Churchland, P. (1985). Reduction, Qualia and the Direct Introspection of Brain States. *Journal of Philosophy*, 82, 8–28.

Churchland, P. (1989). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science.* Cambridge, MA: MIT Press.

Coleman, S. (2009). Why the Ability Hypothesis is Best Forgotten. *Journal of Consciousness Studies*, *16*(2–3), 74–97.

Conee, E. (1994). Phenomenal Knowledge. *Australasian Journal of Philosophy*, *72*, 136–50.

Cresswell, M. J. (1985). *Structured Meanings*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991). Epiphenomenal Qualia? In *Consciousness Explained*. Boston: Little, Brown and Company.

Dennett, D. C. (2005). *Sweet Dreams*. Cambridge, MA: MIT Press.

Hintikka, J. (1975). Impossible possible worlds vindicated. *Journal of Philosophical Logic, 4*(3), 475–484.

Jackson, F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly, 32*(127), 127–136.

Jackson, F. (1986). What Mary Didn't Know. *The Journal of Philosophy, 83*(5), 291–295.

Langer, S. (1951). *Philosophy in a New Key*. Cambridge, MA: Harvard University Press.

Lewis, D. K. (1979). Attitudes De Dicto and De Se. *Philosophical Review, 88*(4), 513–543.

Lewis, D. K. (1983). Postscript to 'Mad Pain and Martian Pain. In David Lewis. *Philosophical Papers Vol. I*. New York: Oxford University Press, 130–32.

Lewis, D. K. (1988). What Experience Teache. In *Proceedings of the Russellian Society*. Sidney: University of Sidney.

Lewis, D. K. (1990). What Experience Teaches. In William Lycan (ed.), *Mind and Cognition*. Oxford: Blackwell, 499–518. (Reprint)

Lokhorst, G.-J. C. (2000). 'Why I am Not a Super Turing Machine'. Manuscript of a talk given at University College London on 24. May 2000. Retrieved in December 2010 from <*http://homepages.ipact.nl/~lokhorst/hypercomputationUCL.pdf*>.

Nemirow, L. (1980). Review of 'Mortal Questions' by Thomas Nagel. *Philosophical Review*, *89*, 473–77.

Nemirow, L. (1990). Physicalism and the Cognitive Role of Acquaintance. In William Lycan (ed.). *Mind and Cognition*. Oxford: Blackwell, 490–99.

Nida-Rümelin, Martine (2009). 'Qualia: The Knowledge Argument'. Published online in Edward Zalta (ed.): *Stanford Encyclopedia of Philosophy.* Retrieved on 12.2.2010 from <*http://plato.stanford.edu/entries/qualia-knowledge/*>.

Penrose, R. (1989). *The Emperor's New Mind*. New York: Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind*. New York: Oxford University Press.

Perry, J. (1977). Frege on Demonstratives. *Philosophical Review*, *86*, 474–497.

Perry, J. (1979). The Problem of the Essential Indexical. *Noûs*, *13*, 3–21.

Quine, W. V. O. (1956). Quantifiers and Propositional Attitudes. *The Journal of Philosophy* LIII (5: March), 177–187.

Rast, E. H. (2007). *Reference and Indexicality*. Berlin: Logos.

Richard, M. (1983). Direct Reference and Ascriptions of Belief. *Journal of Philosophical Logic* (12), 425–452.

Stalnaker, R. (2004). 'Knowing Where We Are, And What It Is Like'. Manuscript of a talk given at NYU's La Pietra Conference on Consciousness, Florence 2004, version of 2006 retrieved from <http://www.nyu.edu/gsas/dept/philo/ faculty/block/lapietra/Stalnaker.pdf>.

von Stechow, A. (1982). *Structured Propositions*. Technical report of the SFB 99. Konstanz: Universität Konstanz.

Thomason, R. (1980). A Model Theory for Propositional Attitudes. *Linguistics and Philosophy* 4, 47–70.

Tye, M. (2009). *Consciousness Revisited*. Cambridge, MA: MIT Press.