So You Think You Exist? — In Defense of Nolipsism

Jenann Ismael
John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721

Abstract

Human beings think of themselves in terms of a privileged non-descriptive designator — a mental "I". Such thoughts are called "*de se*" thoughts. The mind/body problem is the problem of deciding what kind of thing I am, and it can be regarded as arising from the fact that we think of ourselves non-descriptively. Why do we think of ourselves in this way? We investigate the functional role of "I" (and also "here" and "now") in cognition, arguing that the use of such non-descriptive "reflexive" designators is essential for making sophisticated cognition work in a general-purpose cognitive agent. If we were to build a robot capable of similar cognitive tasks as humans, it would have to be equipped with such designators.

Once we understand the functional role of reflexive designators in cognition, we will see that to make cognition work properly, an agent must use a *de se* designator in specific ways in its reasoning. Rather simple arguments based upon how "I" works in reasoning lead to the conclusion that it cannot designate the body or part of the body. If it designates anything, it must be something non-physical. However, for the purpose of making the reasoning work correctly, it makes no difference whether "I" actually designates anything. If we were to build a robot that more or less duplicated human cognition, we would not have to equip it with anything for "I" to designate, and general physicalist inclinations suggest that there would be nothing for "I" to designate in the robot. In particular, it cannot designate the physical contraption. So the robot would believe "I exist", but it would be wrong. Why should we think we are any different?

1. The Mind/Body Problem

I look around and see the world, and when I do I see it from a certain perspective. I see the world as a spatial system with myself located in it, and I see it from the perspective of where I am. My perceptual system locates objects with respect to me. For example, my visual system represents objects in a polar coordinate system with myself at the origin — the focal point. On the basis of my perceptions I make judgements about the way the world is, and adopt goals for changing it. Most of my goals are egocentric — I want to change my own situation in the world. I am equipped for this purpose with various causal powers. I have the ability to perform actions that have effects on my surroundings. These causal powers are centered on my location in the world. I have a body, and I act on the world by moving various parts of my body.

This simple self-description of myself and my place in the world seems uncontroversial, but it leads to perplexing philosophical problems. Although I am intimately connected with my body, and can only act on the world via my body, I do not think of myself as simply *being* my body. When I turn my gaze downwards and see my own body, I think of myself as being "up here looking down". This follows from the way my perceptual system represents the objects I see as being in front of me, with I myself being located at the focal point of my visual field. Anything that I can see is in a different physical location than I am. This includes the parts of my body that I can see, and so they can be neither me nor a part of me. The focal point of my visual field is located inside my head, between my eyes, so I think of myself as being "in here". This leaves open the possibility that I am some physical system that is a proper part of my body and located inside my head — perhaps my brain, or my pineal gland. But it also seems to leave open the possibility that I am something entirely different from my body that is simply residing there in my head. Thus is born the mind/body problem — what kind of thing am I, and what is my relationship to my body?

Familiar philosophical jargon puts this by saying that I am a "self", and then asking what kind of thing selves are. Philosophers have traditionally attacked the mind/body problem by observing that they have various kinds of self-knowledge and then spinning out the consequences of that knowledge. It should be noted that this is the approach that generated the problem in the first two paragraphs. Although we will stop short of rejecting this approach, we will call it into question, and we will entertain the radical solution to the mind/body problem that we call "nolipsism" — there are no selves. Literally, we do not exist. It will be argued that there is more to be

said for this position than might be supposed, although, of course, if it is true then *we* cannot say it.

2. Privileged Access

How might one address the mind/body problem? One venerable strategy has been to focus on the fact that I seem to have privileged access to myself. This is manifested in several different ways. One is Descartes' *cogito* argument. Necessarily, if I have a thought then I exist. Thus if I think that I exist, it follows that I do exist. This is something I cannot be wrong about. Does this show something interesting about selves? It suggests that we can at least be confident that we exist and hence that nolipsism is false. But it will be argued below that this reasoning is fallacious.

Another kind of privileged access is my introspective access to my own mental states. I can tell, in a way that no one else can, that I am having certain thoughts, that the apple on the table looks a certain way to me, or that my finger hurts. The states and events that I introspect are "mental". Presumably there are corresponding physical states and events occurring in my body and causally responsible for my being in the mental states or for the occurrence of mental events. It would be parsimonious to identify the mental states and events with their physical counterparts, but there are familiar arguments to the effect that they are distinct. Jackson's "Mary argument" seems to establish that what I know when I know how red things look to me is distinct from any physical facts about the physical structure of the world.[1] It is tempting to conclude that mental states are not physical states, but that is a non sequitur. All that follows immediately is that the objects of knowledge are different, i.e., mental propositions and concepts are different from propositions and concepts about the physical counterparts of mental states and events.

Token physicalism argues that mental events are the same events as the corresponding physical events, in the same sense that a flash of lightning is the same event as the corresponding electrical discharge. This is based upon a general view about the individuation of events, and we find it convincing. This has the consequence of identifying mental events with physical events, but leaves other kinds of mental objects unexplained. For example, having or feeling a pain is identified with a neurological event, but the pain itself is distinct from the having of the pain — it is not an event. As

---

[1] Frank Jackson, "What Mary didn't know", *Journal of Philosophy* **83** (1986), 291-295.

such, this strategy does not identify the pain with anything physical. The same point can be made about perceptual images, qualia, etc. There does not seem to be anything physical that is even a candidate for being identified with these mental objects. For example, an image cannot be identified with neural activity. The latter is an event, and if it can be identified with anything mental, it must be the having of the image rather than the image itself. Similarly, a pain can recur. Each occurrence of it is a separate mental event, but the pain is something different from any of its occurrences. Our mental lives are desnely populated with such mental objects. We have introspective access to them, and they are apparently not physical. It seems this should tell us something about what sort of thing we are, although it is not clear exactly what conclusion we should draw from this.

The connection between I and my thoughts, percepts, and other mental states and occurrences is perplexing. I "have" my thoughts and percepts. It is tempting to say that they occur "in me". Presumably my having them has physical counterparts occurring within my body. (Note, however, that the counterparts might not occur within that part of the body that is a candidate for being me, i.e., that is located at the focal point of my visual perception.) What is it that makes them *my* thoughts and percepts? It is not just that their physical counterparts occur in my body. It is at least possible that two different persons, with distinct mental lives, could share a body or part of a body. Consider split brain cases, multiple personalities, and perhaps even Siamese twins. So what makes a thought or percept mine? It seems to be a nonphysical fact about it. If this is right, perhaps it should be concluded that I am not a physical thing.

3. *De Se* Representations

The traditional approach to the mind/body problem is to take at face value our internal view of ourselves, and try to find a theory of the relationship between mind and body that accommodates it. Our self-description is accepted uncritically as data. We are going to call this strategy into question, but preparatory to doing this let us to call attention to an important aspect of our self-representation. It is essentially *de se*.

A *de se* representation is one that is expressed with the first-person pronoun "I". The peculiar logic of *de se* representation was brought to the attention of philosophers by a collection of articles by Casteñeda and Perry.[2] We will adapt an example of

[2] H. N. Castañeda, "On the logic of self-knowledge", *Nous* **11** (1976), 9-22; "On the logic of attributions

Perry's to bring out its most important features. Imagine a man, Rudolph Lingens, who finds himself, emerging from a nap, lost and suffering from amnesia in the Stanford Library. He has no beliefs except those he acquires on the basis of his immediate experience. He has no identifying knowledge of himself or his location. His wallet is gone, and there are no signs in sight. He speaks truly when he says "I don't know who or where I am". Suppose, as he wanders the stacks, picking up and flipping through random volumes, he happens on a biography that contains a complete account of his own history. He reads the entire book without an inkling that it is he who is being described. Nothing in the historical account itself, nothing in the objective third-person facts about Rudolph Lingens, tells him that he, himself is that man. He could have a complete account not only of his own life, but of the entire history of the world, beginning to end, and it would give him no clue as to his own identity. It would be as useful to him in his ignorance as the map of a city would be to a lost man who is unable to identify his location on the map. He might even read with interest how Lingens once woke in the Stanford Library in an amnesiac fog, and think to himself, "Poor bloke, I know how he felt". Unless he knows that he, himself is Rudolph, and he, himself is in the Stanford Library, nothing in the objective account of the facts could convey that information. There is nothing that the author could have added, employing only descriptive vocabulary, that would do the trick. Just as the lost man needs for someone to point out his location on the map, Lingens needs a pointer to his identity and location in the world. He is missing a crucial piece of information — information he would express with the exclamation "I am Rudolph Lingens and I am in the Stanford Library". That is not captured in an objective account of the history of the world. It must supplement it.

The crucial observation here is that thoughts formulated using "I" and "here" cannot be reformulated using only descriptions of persons and places. The same thing is true of "now". "I", "here", and "now" are non-descriptive designators. Lingens can know every purely descriptive fact there is to know and still not be able to infer who or where he is or what time it is. We will refer to "I", "here", and "now" as *reflexive* designators.

---

of self-knowledge to others", *Journal of Philosophy* **65** (1968), 439-456; John Perry, "Frege's theory of demonstratives", *Philosophical Review* **86** (1977), 474-497; "The problem of the essential indexical", *Nous* **13** (1979), 3-22. For more recent work by Perry, see *Reference and Reflexivity*, CSLI Publications, Stanford, 2001.

4. Reflexive Designators

Let's list the semantic oddities of the *de se* representation "I":

(i) each person can think of himself himself using "I" without knowing any identifying fact about himself,

(ii) one can possess a complete, objective description of himself, a list of all of one's intrinsic properties and relations to other objects, intrinsically described, without knowing whether "I" applies to it,

(iii) one cannot refer to someone else using "I", no matter how mistaken his self-conception, no matter, even, if everything he believes about himself is true of someone else.

The first two were illustrated in the example of the previous section. We can adapt it to illustrate the third; imagine that Lingens wakes up, not amnesiac, but deluded. Suppose that he wakes up believing that everything Elvis Presley believes of himself is true of him (i.e., Lingens). So he and Elvis have, property for property, identical descriptive self-conceptions, and yet, undeniably, refer to different people when they utter "I".

How this works is a complicated question that requires some delicacy in setting out; the information expressed by Lingens exclamation "I am Lingens, and I am in the Stanford Library" is analogous to that conveyed by the placement of the red dot on a map. The red dot picks out a physical location (in physical space, not on the map) simply by being there. It also indicates a location on the map, and thus coordinates the map with physical space. Similarly, a person's thought refers to a place as *here* simply by virtue of her being there. Her thought refers to a time as *now* simply by being at that time. And she thinks of herself as *I* simply by being that person. These representations secure their designata non-descriptively — simply by virtue of the cognizer's having a location in time, physical space, or the space of persons. In this, they are like the red dot on the map, although *now* and *here* are more like moving dots. They are like the pointer on a GPS (global positioning system) that moves across the map displayed as the GPS moves.[3]

An observation that will be important later is that reflexive designators can designate different kinds of things, and it may not be more than conventionally determinate what they designate. E.g., does the pointer on my GPS designate itself, or the GPS, or

[3] Modern GPS's often contain digital maps and display the location of the GPS on a small LCD screen.

its location, or what? For our use of the GPS, it makes no difference which we say, and we could conventionally stipulate any of these answers. Functional facts about the GPS do not determine a designatum, and they are all that could determine a designatum "objectively". So it is open to us to adopt whatever conventional stipulation we care to adopt, or to leave the matter undetermined, in which case there is really no fact of the matter about what the pointer represents.

5. The Need for *De Se*

The mind / body problem arises from the fact that we think of ourselves in a special non-descriptive way that, by virtue of being non-descriptive, leaves open the question "What am I?" That is, we employ a *de se* designator in our routine cognition. It begins to seem mysterious that we should do this. What is the point of having a *de se* designator at all, particularly if it gets us into such a philosophical muddle? What will be argued is that there are purely computational pressures on the design of a sophisticated cognitive agent that can only be satisfied by providing it with various kinds of reflexive designators, including *de se* designators. Sophisticated cognitive agents literally cannot be made to work in a complex environment unless they are equipped with *de se* designators.

These observations involve an important change of perspective on the mind / body problem. The traditional approach to the mind / body problem takes our internal view of ourselves at face value, and tries to find a theory of the relationship between mind and body that accommodates it. Our self-description is accepted uncritically as data. We are going to approach things in a different way by looking from the outside, in, assuming nothing about selves but that they are designata of *de se* designators, and seeing what can be learned from an examination of the functional role of the designator. What are the conditions under which a being has a need for *de se* designators, and how do they give rise to the problem of understanding the relationship between minds and bodies? This is to adopt the "design stance".

Suppose we want to build a sophisticated cognitive agent — a robot capable of performing intellectual tasks analogous to those performed by human beings. What would this involve? We will assume without argument that a human-like cognitive agent thinks about things in the world in terms of mental representations of them, and that at least some important parts of human rational thought involve manipulating

mental representations. We can think of these mental representations as comprising a system of "mental symbols". Building a cognitive agent involves implementing a system of cognition in an underlying physical structure — a physical (perhaps biological) computer. Our claim will be that the need for reflexive designators in general, and *de se* designators in particular, arises from the demands of practical reasoning in a cognitive agent capable of functioning in a complex and unpredictable environment. We assume that practical reasoning consists of: (1) the adoption of goals as the objects of some kind of conative state that we will noncommitally call "valuing"; (2) epistemic reasoning about how to achieve goals; and (3) the selection and execution of courses of action discovered in (2). Rather simple considerations give rise to the need for a mental *here* and *now*, and increasing complexity gives rise to the need for *de se* designators.

5.1 *Now* in Epistemic Reasoning

Perception is only possible in a changing world because, after all, perception changes the agent. Truth in such a world must accordingly be indexed to times, and a cognitive agent that possesses knowledge of the way the world is at different times needs a way of indexing its beliefs to times. One way to do this — the human way — is to include designators for times in the agent's system of mental representations.

It will be useful to contrast various kinds of cognitive agents with a chess-playing computer. The latter could be implemented as a simple agent that plays chess by reasoning about what to do. (Real chess computers don't work this way, but any real chess program could be re-implemented within OSCAR[4] so that the agent uses the same search algorithms but reasons about what moves to make.) The importance of this example is that, as we will argue, the chess agent is able to engage in practical reasoning while making only minimal use of reflexive designators. If we are to explain reflexive designators in human cognition as arising from the computational needs of human practical reasoning, we must explain how human practical reasoning differs from that of the chess agent, and how that difference gives rise to the need for reflexive designators.

At first blush it seems that the simplest version of the chess agent does not need a mechanism for temporal indexing, because it does not store beliefs about other times. Its beliefs are only about the current state of the chess board. However, if it is to choose its moves on the basis of practical reasoning, then it must be able to conceive of

---

[4] OSCAR is the artificial rational agent constructed by John Pollock and described in *Cognitive Carpentry*, Cambridge Mass: MIT Press,1995.

the board having one state at the present time and a different state at some future time. First, its goal (e.g., black wins) is about the future. That is, the goal is that *there will be* a board position of a certain sort (a winning position for black). To plan for the achievement of those goals, the agent has to have beliefs to the effect that different kinds of moves will have specific effects on the board position, i.e., that if the board is initially in a certain position it will subsequently be in another position. This requires being able to distinguish between board positions occupied at different times. However, it does not require the ability to actually think about the times themselves. It requires no more than a temporal ordering of positions. The agent needs a way of representing what comes before what, but this does not require designators for times.

A natural need for temporal designators does not seem to arise until the agent begins to form beliefs about the physics of its environment. Then it needs a way of representing temporal duration. This seems to require temporal designators, i.e., the ability to think about times rather than just the passage of time.

The need for the reflexive mental designator *now* arises from more sophisticated cognitive or computational pressures. *Now* refers to the *current time*. For the chess computer to reason about how to achieve goals, it must be able to distinguish between its current board position and possible future board positions. This by itself does not require a designator for the current time. The belief of the chess agent could instead use a tensed copula, giving it the form "The position is *B*" (as opposed to "The position is *B* at the present time". The tensed copula relieves the agent of the need for a representation of the current time.[5] Given the tensed copula and temporal reference, we can define the reflexive representation *now* as "the time it is". But if the agent has temporal ordering and the tensed copula, without temporal reference, we cannot define a temporal designator for the current time. A reflexive temporal designator can only be introduced when the agent has temporal representations in general, and the latter only seem to be necessary for the agent to have rather sophisticated physical knowledge of how the world works.

For practical reasoning, the agent must be able to distinguish between the current state of the world and possible future states. This requires at least the tensed copula. The tensed copula can be defined in terms of *now*, viz., "P is true" means "P is true at the present time". So if the agent has temporal representations in general, then the

[5] It is perhaps worth noting that the English word "now" is actually an adverb, not a pronoun. "now" means "at the current time". This relates it closely to the tensed copula. It is unclear whether this observation is of importance.

tensed copula and *now* are interdefinable. It is worth noticing that neither can be defined "descriptively", as "the time that satisfies description *D*". If that were to work, description *D* would have to be a different description for each instant of time, and so there would be no general description that could do the job. Thus very general cognitive pressures require the agent to have some way of thinking non-descriptively of the present time.

5.2 *Here* in Epistemic Reasoning

Perception provides humans with an egocentric view of the world. Visual, tactual, proprioceptive, and perhaps some other modes of perception have a "perspective", and the human agent has a position in space relative to that perspective. Roughly, we perceive the world from where we are. We can imagine agents that differ from us in this respect. For example, the chess agent has input (from the keyboard) that updates its knowledge of the board positions in the game it is playing. But its knowledge is about "the game", "the board", etc. As it is only aware of one game, board, etc., there doesn't seem to be a need for reflexive designators for spatial location. We can similarly imagine an artificial agent with "distributed sensors" that have fixed positions in the world. For example, the agent might reside in a room with video cameras mounted in each corner of the ceiling. The information derived from perception (via the video cameras) would still give perceived objects spatial locations, and that requires a coordinate system, but that coordinate system might be shared by several similar agents all residing in the same room and having physical implementations in different bodies.

Vision

In contrast, human visual perception is perspectival, providing knowledge of objects relative to an egocentric coordinate system. Roughly, this is a polar coordinate system with the agent at the origin. The beliefs the agent acquires via perception are beliefs about what is going on at particular spatial locations identified with reference to this perceptual coordinate system. The beliefs actually make reference to locations in the coordinate system. This requires a way of representing the locations, and hence of representing the coordinate system itself. One way of picking out a coordinate system is relative to the locations of some specific objects, however we cannot form beliefs about objects until we perceive some objects, and that involves a prior ability to form beliefs about locations in our perceptual coordinate system. So the perceptual coordinate system cannot be anchored conceptually by reference to objects in the world. We must be able to represent locations in this coordinate system before we can form beliefs

about objects in the world. The only way to do this is to have a designator *here* that designates the location of the origin of the coordinate system, and a designator *before* (*in front of here*) indicating a direction from *here*. We also need designators like *up*, and *right* or *left* indicating orientation.

The designators *here*, *before*, and *up* cannot get their content from descriptions relating them to objects in the world, because we must be able to employ these designators prior to acquiring perceptual knowledge of objects in the world. Given a *de se* designator, we might try defining *here* as "where I am now", *before* as "in front of me", and *up* as "in an upward direction relative to me". The first definition is plausible, but the others are not. "In front of me" and "in an upward direction relative to me" already presuppose the directionality and orientation relative to *here* that is being defined.

If we are building an agent, and it is only intended to function in a narrowly circumscribed environment whose general properties we know, we might give the agent built-in knowledge of that environment (an "a-priori world model"), including built-in knowledge of its own body. This would make it possible to have a description that picks out the agent's body uniquely, and then we could design the agent's cognitive architecture in such a way that perception gives it beliefs about states of the world located relative to its body (designated descriptively). In this case, *here*, *before*, and *up* can be descriptive designators constructed in terms of a descriptive designator designating the agent's body. Notice, however, that the descriptive designators we choose must play a privileged role in the agent's epistemic norms. The agent's epistemic norms must automatically locate perceived objects relative to the object (body) described. That is necessary for the agent to be able to acquire knowledge of its surroundings simply on the basis of perception. Thus we cannot require the agent to *discover* where, in its visual field, is the object (body) described. If the agent had to do that before judging where perceived objects are, it would not be able to get started. In this sense, the descriptive designator we choose for picking out the body isn't really functioning descriptively.

The biggest problem with designing an agent in this way is that it is "brittle" in the sense that it will not be able to function in an environment that differs in any way from its built-in world model. The agent will have to judge that perceived objects are located in proximity to the object described by the privileged designator even when things go wrong and nothing fits the description. If the agent subsequently discovers that nothing fits the description, that will defeat all of its earlier perceptual judgments and all of its putative contingent knowledge of the world will evaporate.

If an agent must be able to function in a wide variety of environments with rather

unpredictable properties, the use of a descriptive designator is not an option. A "flexible" agent needs the designators *here*, *before*, and *up* as anchors for the coordinate system used by perception, and these designators cannot be descriptive. They must be primitive elements of the agent's cognitive architecture. Objects in the world are represented as having locations picked out by descriptive designators defined in terms of these reflexive designators rather than the reflexive designators getting their content from some objects in the world. The reflexive designators just act as anchors for relating different perceived objects. Once an agent has a fair amount of knowledge of the world, it can ask where *here* is, and answer this with respect to its body or some other interesting objects, but cognition must begin by employing *here*, *before*, and *up* as primitive designators.

So in "flexible" agents, visual perception provides information about *here*, *before*, *up*, and also *now*. *Here*, *before*, and *up* generate a three-dimensional spatial coordinate system, and if *now* is supplemented with temporal reference we get a four-dimensional coordinate system. (Note that for temporal reasoning we need temporal directionality, i.e., past and future directions of time, just as we need *before* and *up* for spatial reasoning.)

Touch

Touch (haptic perception), like vision, is perspectival, locating objects in a polar coordinate system whose origin is centered on the body. However, at least in humans, the origin of the tactual coordinate system is not in the same place as the origin of the visual coordinate system. Introspectively, the origin of the tactual coordinate system is located somewhere on the body below the head. Furthermore, the *before* and *up* dimensions of the tactual coordinate system are often oriented differently from those for the visual coordinate system. For instance, if I am looking over my shoulder, what is before me visually is behind me tactually. And if I am looking between my legs, what is up visually is down tactually. This indicates that there are distinct visual and tactual *here*'s, *before*'s, and *up*'s.

Although vision and touch provide information about the world via separate coordinate systems, we regard the objects perceived tactually to be in the same physical space (and usually to be the same objects) as those perceived visually. It has often been noted that it is a contingent fact that vision and touch give us knowledge of the same physical space. Presumably we could build an agent that had to discover this fact by a combination of induction and inference to the best explanation. For most agents this is a completely predictable aspect of their environnment and so it makes cognition more efficient to simply build this into the agent's cognitive architecture. However, this is more difficult to achieve than might be supposed. The source of the

difficulty is the observation that the visual and tactual coordinate systems are different and not even stably correlated. To get a stable correlation we must at least take account of proprioception, which provides information about how the visual and tactual sensors are oriented with respect to each other. Presumably, with this added information, we can build into the agent's cognitive architecture the expectation that vision and touch provide information about a common space and (generally) common objects. Of course, there are visual objects like shadows, rainbows, and holograms that have no tactual correlates, and there are tactual objects like wind or objects felt in the dark that may lack visual correlates, so all of this must be rather complicated. However, we will not pursue the details here.

It is worth noting that although vision and touch are perspectival, locating objects in a polar coordinate system with the agent at the origin, when we think about the physical world abstractly we think of it in terms of a fixed three-dimensional space and we think of ourselves as moving around in it, rather than thinking about it in terms of one of our perceptual coordinate systems.

5.3 *De Se* Goals

An agent only capable of epistemic cognition about the physical world around it does not seem to have need for a way of thinking of itself. This is particularly obvious if it is just an idle spectator rather than a causal force on its environment. So although agents having moderate epistemic sophistication need the reflexive designators *here* and *now*, they do not need *de se* designators. When do *de se* designators become necessary? Our suggestion will be that the need for *de se* designators arises in part from the goal structure of sophisticated practical reasoners and in part from the need to reason about how to achieve goals.

Human goals tend to be personal (although not exclusively so). Goals derive from what the agent values, and valuing is egocentric in humans. Humans tend to value states of affairs in which they themselves play a particular role. If there were a description (e.g., "the first type 17 robot constructed") that is guaranteed to pick out the agent in any world it is apt to be in, its conative machinery could generate valuings of states of affairs involving that description rather than a *de se* designator, and the resulting goals would be guaranteed to be "about" the agent itself. If the agent also had knowledge (perhaps built-in) about how to achieve such goals, then it could engage in full-fledged practical reasoning without having *de se* designators. However, for general-purpose agents operating in unpredictable environments, or extremely variable environments, there will be no such description. The only way to formulate personal goals for such

agents is by using a non-descriptive designator.

Humans have many different kinds of goals. I have low-level goals such as the alleviation of *my* hunger, but also high-level goals concerning such things as *my* country, *my* as yet unborn children, the books *I* will write over the next twenty years, *my* personal appearance, *my* knowledge of astrophysics, *my* summer vacation, etc. Even a goal like world peace is really egocentric. What I value is peace in *my* world among beings like *me*. These goals can only be formulated using a *de se* designator. Agents capable of having such goals must be constructed so that their conative machinery produces valuings of *de se* states of affairs, i.e., produces *de se* goals.

Our conclusion is that *de se* goals are essential in agents that (1) have wide-ranging personal goals (i.e., goals in which they themselves figure in a privileged way), and (2) their operating conditions are sufficiently unpredictable to make it impossible for either evolution or their designer to seize upon a descriptive designator beforehand and build that into their conative and cognitive machinery.

5.4 Knowing about Actions and their Effects

It does no good to have goals unless the agent can figure out how to achieve them. In order to reason about how to achieve a goal, an agent must make judgments about what actions it can perform and what their likely effects are. These ability-judgments are *de se* — in practical reasoning, what is at issue is what *I* can do.

There is a difference between doing something on purpose (intentionally) and doing it accidentally or having it simply happen to you. For instance, there is a difference between your moving your arm and your arm moving without your willing it. For practical reasoning, we want to know what we can do intentionally and what is apt to happen if we do. A simple agent might have this knowledge built into it, but a more sophisticated agent must be able to acquire new knowledge about what it can do as its skills and physical capabilities change. It seems that the judgment that I will be able to do something in certain circumstances is generally based inductively on the observation that I often have done it in those circumstances. This requires me to have the ability to tell (not necessarily infallibly) that I am doing or trying to do something intentionally (e.g., moving my arm) rather than its just happening to me without my initiating it. It seems that this is something humans can introspect — we can tell what we are trying to do. It is hard to see what other alternative there could be.[6] It seems that the cognitive

---

[6] G. E. M. Anscombe addressed thisquestion at length in *Intention* (Basil Blackwell 1957). But in the end she did not produce an account of how such self-knowledge is possible. Her conclusion was simply

architecture of a practical reasoner must contain machinery for introspecting what one is trying to do. The output of such an introspection module will be *de se* — *I* am trying to do such-and-such. Note that a properly equipped agent may be able to make such judgments without knowing what it is to do something intentionally. It certainly need not have at its disposal any kind of philosophical analysis of intentional action. It might not even have the "in principle" ability to find such an analysis. That would not hamper its ability to engage in practical reasoning. All that practical reasoning requires is that the agent makes such judgments and uses them in deciding what to do.

We first generated the need for *de se* representations by looking at egocentric goals and noting that they must be *de se*. It is hard to imagine how we could have an agent none of whose goals are egocentric. But it is worth noting that even if an agent's goals were not egocentric it would still need *de se* representations to reason about what it can do intentionally. So this constitutes a separate source for the need for *de se* representations.

5.5 Reasoning about How to Achieve *De Se* Goals

Reasoning about how to achieve goals requires more than judgments about what we can do. It also requires judgments about what is apt to happen if we do those things. If the goals are *de se*, this requires the agent to engage in epistemic reasoning about *de se* propositions. How is that possible?

I possess several different kinds of *de se* goals. Some are about my inner states — e.g., the alleviation of my hunger or pain. Others are about my body — I want to get a haircut. Still others are not directly about my body but are about things causally connected to my body — I want my children to get a good education. Most involve a mixture — I want to attend a chamber music festival, I want to read a new novel by my favorite author, I want to dine with friends at a new restaurant.

Consider my goal of alleviating my hunger. To achieve this goal, I must learn that my ingesting certain substances will usually be followed by diminished hunger. Furthermore, I must learn that I can ingest such substances by intentionally moving my body in certain ways under specified circumstances. Both of these facts that I must learn are *de se*. To learn that my ingesting certain substances will usually be followed by diminished hunger, I must be able to tell that it is *I* that is doing the ingesting. To do this I must be able to pick out my own body in the world. Similarly, to learn that I

---

that part of what it is to do something intentionally is to know that you are are. She gave no explanation for how you can know that.

can move my body in certain ways, I must be able to tell that it is *my* body that is moving.

I do things by moving my body or parts of my body in various ways. So to reason about the effects of my actions, I must be able to identify my own body. It is interesting that that does not seem to require me to be able to locate myself (as opposed to my body) except insofar as I am at the same location as my body. Human beings are aided in locating their bodies by the fact that the point from which they see is located on their body, and the movements of their limbs are generally readily apparent perceptually. This makes it convenient for humans to be built so that they regard themselves as being at the focal point of visual perception, and to think of themselves as having physical extremities projecting outwards from that location and enabling them to act upon the world. However, we can imagine cognitive agents in which these matters are not so nicely organized. Consider a "distributed" agent that is confined to a single room and whose perception is provided by video cameras permanently mounted in the corners of the room. The visual field of such a system need not encode information in a polar coordinate system. It can use a straight-forward three-dimensional coordinate system of the sort that humans use to represent physical space. Suppose the seat of cognition for this agent is a box of electronics mounted on the ceiling, and those electronics remotely control robot hands mounted on little electric carts that run around on the floor. This agent will still have *de se* goals and need *de se* beliefs about what it can do, and for this purpose it will need beliefs about the locations of its hands. As the robot hands are able to move around the room and carry out physical tasks, there will be no way to assign them a fixed location in the agent's visual field (its visual representation of the world). How might this agent acquire the kind of *de se* knowledge about its own actions that is required for achieving *de se* goals? One way to do this would be to let proprioception provide the agent with *de se* knowledge (it certainly does in humans). If the agent can tell proprioceptively when it is moving in certain ways, then it could correlate its movements with the movements of a specific body in its visual field, and then inference to the best explanation might lead it to conclude that the movements of that body are *its* movements. This requires that proprioception provides information about bodily movements in a form that enables the agent to identify them with visually perceived bodily movements. Proprioception must yield more than just knowledge of what the movements feel like. It must yield spatial characterizations of a sort that can be compared with the spatial characterizations generated by vision.

As long as the robot hands can be readily perceived, and the agent can sense its hand movements proprioceptively, it can discover inductively which hands it can

control. At least in principle, it can then learn inductively that it can alleviate its "hunger" by backing its robot hands up to a wall socket where their batteries are recharged. The point of this example is that such an agent can engage in practical reasoning about how to achieve *de se* goals without vision providing it with any *de se* beliefs. Introspection must provide *de se* beliefs to the effect that the agent is hungry, and proprioception must provide *de se* beliefs about what it is doing, but vision need not.

If vision does not produce *de se* beliefs, then it provides no direct basis for the agent to make a judgment about where it is. But the distributed agent has no need for such a judgment. All it must be able to determine is where its robot hands are, and that is something it does inductively by discovering which hands it can control. We might be tempted to insist that although the agent does not judge itself to have a location, it nevertheless does. Its location is the distributed location consisting of the locations of all of its robot hands. But why should we say that? What about the seat of cognition mounted on the ceiling? Should that also be counted as part of the location of the agent? If that is to be counted, how about the city power plant that produces the electricity used by the seat of cognition? There does not seem to be any clear line to be drawn between what counts as part of the agent and what counts as facilities supporting the operation of the agent.

It is not clear that the distributed agent actually has a location. This is because it has no need to reason about its own location. In this respect, it is quite unlike human beings. We take perception to locate perceived objects with respect to ourselves, and so conversely we are located in a certain place relative to the objects we perceive. A partial explanation for why we are so constructed is that, unlike the distributed agent, our sensors move around in the world and so cannot, without further inference, locate perceived objects in a fixed three-dimensional reference frame. Because our sensors move, our view of the world must be perspectival. However, at least in principle such perspectival judgments need only locate objects with respect to *here*, not necessarily with respect to *me*. This suggests that it is only a matter of convenience that human perception locates objects relative to the self.

The general lesson in all of this is that in order for an agent to be able to reason about how to achieve *de se* goals, it must have epistemic norms enabling it to identify its own actions in the coordinate system used by its perceptual system. Human epistemic norms do this in part by locating the self at the origin of the visual coordinate system and locating the body (the locus of actions) in close proximity to the self. But it appears that this is just one way to solve the problem. There could be agents that did

not locate themselves in physical space at all, and they would still be able to reason about how to achieve *de se* goals.

5.6 *De Se* Memories

The next thing to observe is that to engage in practical reasoning about the achievement of egocentric goals, a cognitive agent must have beliefs that are about itself at different times. First, egocentric goals are about the agent's future situation. To reason about how to achieve them, the agent must form beliefs about what will be true of it in the future if it does various things. These are *de se* beliefs about the future. Second, the agent must reason inductively about what it can do and what the effects of its actions are likely to be. This requires monitoring what one did in the past and what happened to oneself as a result. These are *de se* beliefs about the past. Presumably, our *de se* beliefs about the future are based inductively on such *de se* beliefs about the past.

An agent's access to the past is ultimately via memory. Any other source of historical knowledge, such as the testimony of others, must be validated by appeal to either memory or previously validated sources of historical knowledge. Memory is fallible, just as is the evidence of our senses, but it must provide us with at least defeasible justification for believing what we seem to remember. Otherwise, we would have no access to the past at all.

De se knowledge of our own past states must derive ultimately from *de se* memories. We can also have purely descriptive memories in which we judge that we were one of the characters described, but the latter identity (that we are the person described) is just further *de se* historical knowledge. As a matter of logic we cannot infer *de se* conclusions from a set of purely descriptive beliefs. So if we are to have *de se* historical knowledge, some of it must come in the form of *de se* memories, and we must treat the latter as giving us defeasible justification for believing their pronouncements.[7]

De se memories make it possible for the agent to reidentify itself over time. It can know that *it* is the one that did so-and-so because it remembers doing it. An agent's *self* is the designatum of its *de se* designator. If we encounter a novel kind of agent, like the distributed agent of section 5.5, and we want to know what its *de se* designator designates, we must take its own pronouncements about its self-identity seriously. We have no other access to what it is thinking about. For example, our initial inclination may be to identify the distributed agent (the robot's self) with the set of its robot

---

[7] For a more extended argument to this effect, see John Pollock and Joseph Cruz, *Contemporary Theories of Knowledge*, 2nd edition, Lanham, Maryland: Rowman and Littlefield, 2000.

hands. But suppose the robot hands are removed from the room each night for maintenance, and new hands left in their place. Suppose the robot tells us that it gets new hands each night, but they all work alike — enabling it to plan ahead for the achievement of long term goals that may require the use of its hands over a period of several consecutive days. We ask how it knows this, and it replies that it remembers this happening every night of its "life", and it remembers formulating such long term goals and pursuing them over a period of several days. If we grant that the robot's *de se* designator does designate something, and we are trying to understand what that is, we have nothing to go on except its reports of its own persistence. If a hypothesis about the robot's self-identity conflicts with our only access to the persistence of the robot, i.e., with its reports of its own persistence, then the hypothesis cannot possibly be warranted. So we would have to reject the claim that the robot's self is identical with the collection of robot hands.

5.7 Reflexive Designators and Computational Pressures

The general conclusion to be drawn from this section is that purely computational pressures deriving from the requirements of situated cognition in a rational agent give rise to the need for the reflexive designators *now*, *here*, and *I*. Purely epistemic considerations require *now* and *here* in agents operating in complex environments. The need for *de se* goals derives from the requirements of practical reasoning in agents with widely varying personal goals, and the need for *de se* beliefs derives from the need to be able to reason about how to achieve *de se* goals. *De se* beliefs are also needed for reasoning about what the agent can do in attempting to achieve goals, and for reasoning about past and future states of the agent.

To better understand the nature of these conclusions, let us make a tripartite distinction. First, we can distinguish between the mental states involved in thought (propositional attitudes) and their propositional objects. Let us take propositions to be the "logical" objects of thought, and give them however much structure that requires. In particular, they may contain various kinds of designators designating individual objects. Thus we do not think of propositions as being sets of possible worlds. What we can noncomittally call "sentences" in our system of mental representation "express" propositions. Propositional attitudes consist of *believing-true, hoping-true, fearing-true,* etc., propositions. They do this by employing mental sentences in various ways.

The manipulation of mental representations is implemented in a physical computational system — a physical (perhaps biological) computer. Computers can be described at various levels of abstraction, and at some levels it is appropriate to talk about

"virtual machines" manipulating symbols. For example, we might write a LISP program that manipulates lists of numerals. Let us call these computer symbols *c-symbols*. We may be begging some questions here against certain construals of connnectionism, but it is our conviction that connectionism is best viewed as a theory about lower levels of implementation and a connectionist architecture that correctly models human cognition must make room for a high level description in terms of c-symbols.

Thus we are led to a tripartite distinction between mental representations, c-symbols, and propositions and their constituents. When we have *de se* thoughts, the propositions we entertain contain logical items we can call *de se designators*, and our mental sentences contain "syntactical" items we can call *de se representations*. There will also be *de se c-representations*, which are just computer symbols used in the implementation of *de se* thought. We don't mean to make any metaphysical hay out of these distinctions. We just want it to be clear whether we are talking about mental items, computational items, or the propositions and propositional constituents they represent.

If we are to build an agent, we do that by implementing the cognitive architecture in a physical computational system. In effect, we program a computer that is connected to the world in various ways. What we have described as the computational pressures giving rise to reflexive designators are really remarks about how to program such a computer to enable it to carry out various tasks. What the computational pressures require most directly is dedicated c-symbols that are treated in special ways during cognitive processing. These c-symbols "correspond to" reflexive mental representations and reflexive designators in the corresponding propositions. However, what is needed to implement epistemic and practical reasoning is the c-symbols. The mental representations and propositional designators are there (if they really are) just because of the c-symbols. And what is needed vis-a-vis the c-symbols is that they play a purely computational role in the implemented cognition. The c-sentences generated by the agent's conative and perceptual systems must contain the reflexive c-symbols and the computational processes must make use of that to mesh the outputs of the systems properly, in effect enabling the agent to form goals and acquire c-beliefs about how to achieve them.

6. What Am I?

Now let us return to the mind/body problem. The problem arises from the fact that we think of ourselves in a non-descriptive (*de se*) way. I am whatever my *de se*

designator designates. Because my *de se* designator is non-descriptive, it is not transparent what kind of thing it designates. How then can we find out what we are?

It is plausible to suppose that the referent of any mental term is determined by its functional role in thought together with the way in which the agent's body is situated in the world.[8] The latter allows the agent's causal connections to the world to play a role in determining reference. This is a general remark about the contents of a cognizer's thoughts. Applying it to *de se* representations, it follows that the referent of my *de se* representations has to be determined by my built-in rules for reasoning with *de se* representations together, perhaps, with facts about how my body is situated in the world. If there is a fact of the matter about what kind of thing I am, it must follow from these computational and causal facts about my cognitive system. This is the *determinate reference principle.*

Suppose we build a sophisticated robot. To enable it to engage in sophisticated practical cognition, we must equip it with a *de se* c-symbol, thus enabling it to think (or at least c-think) of itself in a *de se* way. It then becomes an open question what the robot's *de se* c-symbol represents. Just as for human beings, there is a potential distinction between the robot's body and its self (the object of its *de se* c-thoughts). They may be the same thing, but that remains to be determined. If the robot is sufficiently intelligent, it may become very interested in this question. However, in building the robot, there is no need for us to equip it with the resources for answering the question "What am I?" *directly*. The robot will be able to perform its routine cognitive tasks entirely adequately without knowing the solution to the mind/body problem. If there are facts about the robot's cognition that determine the referent of its *de se* representations, and the robot is sufficiently intelligent, then it can in principle solve the mind/body problem. On the other hand, if there are no facts about the robot's cognitive architecture that determine a solution to the mind/body problem, it follows from the determinate reference principle that there is no fact of the matter about what its *de se* representations represent. If there is nothing such that it is a fact that the robot's *de se* representations represent that thing, then there is nothing that they represent. And again, that need be no obstacle to a robot's performing its routine cognitive tasks or getting around in the world. It is a useful fiction for the robot to c-think that "it" exists, but there is no need for that to be true. If the robot c-thinks "I exist" without there being anything that "I" designates, then there is nothing that actually *thinks* "I exist". The robot just c-thinks

---

[8] For a detailed account, see chapter five of John Pollock's *How to Build a Person*, MIT Press, 1989.

there is. Of course, the robot's body exists, but that need not be designated by the robot's *de se* c-symbol.

It is clear that the cognitive architecture of an agent with a *de se* representation need not determine its referent "directly", i.e., there is no need for a simple rule built into the agent's cognitive architecture enabling it to immediately conclude "I am my body" or "I am a non-physical being" or "I am a supervenient object, supervening on my body by virtue of my body's computational organization". But it is too quick to conclude that because the agent is not equipped with a simple rule of this sort, there is no answer to the question "What am I?" that is forthcoming from some more complex argument employing general inference schemes that serve the agent elsewhere. In fact, searching for such arguments is exactly the business the philosopher of mind is in.

In evaluating standard philosophical arguments that purport to answer the question "What am I?", it will be useful to consider how uncompelling they are when applied to our robot. Because we are antecedently convinced that we exist, we find such arguments more compelling when we view them from the inside as applied to ourselves than we do when we apply them to a robot.

What kinds of arguments are there that purport to determine the referent of a *de se* designator? Let us rehearse a few familiar ones. The most obvious is an abductive argument alleging that the simplest explanation for what we know about ourselves is that we are identical with our bodies. Here we take it for granted that we exist and that our mental states are determined by the physical states of our body. Given this data, it is explanatory to hypothesize that I am identical with my body.

A view insisting that there is nothing non-physical in the world and hence that we must be the most convenient physical thing associated with our activities, viz., our body, is certainly a simple view. The trouble is, it really doesn't explain everything that we think we know about ourselves. For example, most people believe either that when they die they cease to exist even if their body continues to exist for a while, or that they can continue to exist even if their body is destroyed. In either case it follows that they are not their body. However, it is a little hard to see how to defend either of the premises on which this argument turns.

Another familiar argument from the philosophical literature involves brain transplants.[9] If my brain is transplanted to another body, it is tempting to suppose that I

---

[9] See Sydney Shoemaker, *Self-Knowledge and Self-Identity*, Ithaca: Cornell University Press, 1963.

will go with it, in which case I am not identical to my (whole) body. Note that this argument seems to turn on the presupposition that our *de se* memories will go with our brain. It was remarked above that general computational considerations require that we reidentify ourselves by appeal to those *de se* memories. However, for the same reason, intuitions regarding brain transplants are not robust. For example, we can imagine a professional football player who learns that he has an inoperable brain tumor. It is not out of the question that he would opt for a brain transplant so that he can continue to play football, particularly if he is told we can do a core dump of his memories and personality traits to a computer and upload them to his new brain after the operation. Note that the pull of this example also turns in part on the observation that we reidentify ourselves across time by appealing to *de se* memories. By restoring the football player's memories in his new brain, we ensure that, *as he remembers it*, he is still the same football player.

A variant of the brain transplant argument that seems stronger is a kind of *Ship of Theseus* argument. Suppose that at some time in the future many medical procedures are performed like some car repairs are now performed. Doctors maintain a respository of body parts, and if I injure my arm they simply remove it and replace it with another arm. They repair my damaged arm at their leisure and put it into cold storage to be used for another patient. This might not work with brains, but presumably it would work with most other organs. We can imagine that over time Jones and I, both of whom are accident prone, end up purely by chance exchanging all of our major body parts. The body I then have has a stronger claim to being the same body as the one Jones used to have than it does to being the same body I used to have, but this does not tempt me to conclude that I am really Jones. So it seems doubtful that I am the same thing as my body.

Perhaps a more plausible view would be that I am some part of my body, perhaps my brain or some still smaller seat of cognition. This seems a bit ad hoc, but if there are neurological parts of my body that could not be destroyed without destroying me, this at least avoids the preceding argument.

However, we began this paper with a different argument to the effect that I am not my body. We are now in a position to construct a variant of that argument that is more compelling that the original version may have seemed. Human beings locate perceived objects spatially with respect to themselves. That has the converse effect of locating them with respect to visually perceived objects. In human beings, the location of the self relative to perceived objects is made possible by locating the self at the focal point of the visual field. It is not computationally necessary to do that. There is nothing

obviously wrong with building an agent whose cognitive architecture resulted in its locating itself six inches to the left of that focal point. We find that perverse when we try to imagine it, but that is because our cognitive architecture enforces the identification of our location with the focal point. But the only reason for having an agent locate itself in space is to provide a convenient reference point for use in relating perceived objects to one another. Different kinds of agent architectures could work just as well for this purpose.

Human beings have their eyes embedded in the fronts of their heads, and accordingly they locate themselves somewhere inside their heads. That is where it appears to them visually that they are. But imagine a somewhat different kind of creature whose eyes were mounted on the ends of willowy stalks extending outwards some distance from the head. The focal point of the visual field of such an agent might be three feet in front of its head, and it would be natural to construct the cognitive architecture of such an agent so that it took itself to be located at that focal point. The interesting thing about this example is that there need be nothing physical that is at that location. So if the self is genuinely there, then it isn't anything physical.

We can get the same conclusion by imagining a human being with a malformed head that has a big empty space in the middle of it. Suppose that just happens to be the location of the focal point of that person's visual field. Then for her too, there isn't anything physical where she thinks she is. So identifying the self with a physical part of the body does not explain important beliefs that the person has about herself.

Could we insist that the agent is just wrong about where she is? The only access we have to the designatum of the agent's *de se* designator is her beliefs about herself. We have noted that, as a human being, it is an essential part of her cognitive architecture that she believes she is where she seems to be vis-a-vis her visual field. The agent's epistemic access to the world is via beliefs like "There is an apple on the table before me". The agent cannot forsake her belief about her location with respect to her visual field without giving up such beliefs as this, and giving up all of these beliefs would undercut all of her contingent knowledge of the world. The agent will then be left without any objective information she could use to try to locate herself somewhere else.

At this point, it is useful to consider the distributed agent again. The distributed agent need not have any beliefs about where it is. This is because its visual perception is non-perspectival. One might say that the distributed agent is wherever its robot hands are, but that cannot be right if the robot changes hands every night. We might instead suppose that the agent is where it center of cognition is, viz., in the box on the

ceiling, and its effectors are the radios that send out signals controlling the robot hands. Given that the agent has no beliefs about its own location, it is hard to see what could decide this question. In fact, if we ask it where it is, the robot will say, "I don't know what you mean. Physical location is not applicable to me." It seems to us most reasonable to just deny that the distributed agent has a physical location. It is more like a deity that views the room from outside that coordinate system and directly manipulates events in the room. If the agent has no physical location, then of course it is not anything physical.

These days, most of us are physicalists and believe that there is nothing non-physical in the world. But faced with arguments like the above, some philosophers have been tempted to bite the bullet and conclude that the self is non-physical. They have then been faced with the task of explaining what kind of a non-physical thing they might be. We do not feel that convincing answers to this question have been given. Still, one might be convinced that even without an account of what non-physical selves are like, we are forced to conclude that that is what we are.

We don't think so. Let us return to the design stance. If the argument that we are something non-physical is compelling when applied to human beings, it should be equally compelling when applied to robots with the same cognitive architecture. Suppose we want to build a robot that is capable of cognition of human-like sophistication. So we build a physical computational system, and provide it with a cognitive architecture by suitable programming. All we put into the robot was a bunch of physical stuff. By virtue of the way we programmed it, we provided it with a *de se* c-symbol, but we didn't provide it with anything for the *de se* c-symbol to represent. There is no need for that in order to get the robot's cognition to work properly. A *de se* c-symbol is required as an anchor-point for tying various aspects of cognition together. It enables c-thoughts about perception, conative states, and intentional action to interact with each other in the ways required for sophisticated practical cognition. But for that purpose, it makes no difference at all whether there is anything that the *de se* c-symbol represents. And in building our robot, we have not built in anything for the *de se* c-symbol to represent, so there seems to be no reason at all to think that somehow a shadowy non-physical self sneaked in. Such a hypothesis has no explanatory power. We can explain everything there is to explain about how the robot works without recourse to its having a non-physical self. Positing non-physical selves seems tantamount to positing a ghost in the machine. There is no more reason for thinking there is a ghost in my robot than there is for thinking there is a ghost in my attic.

The preceding considerations reflect the fact that the abductive argument is com-

pletely different when applied in the first-person to ourselves and when applied in the third-person to the robot. In applying it to ourselves we take it for granted that we exist and that our mental states are determined by the physical state of our body. Given that data, it would be explanatory to identify myself with some suitable physical structure. But in the case of the robot, the existence of the self is part of what is at issue. It is not part of the data to be explained. The data we have regarding the robot concern its physical constitution and its behavior. Identifying the robot's self with some physical structure is completely unexplanatory. The robot's physical constitution is what it is because we built the robot that way, and the robot behaves as it does because we programmed it to manipulate c-symbols in the way required for sophisticated cognition. To hypothesize a robot self and real thoughts (as opposed to c-thoughts) is completely gratuitous. It buys us nothing. So it is scientifically disreputable to suppose the robot really has a self.

Suppose we all agree that when we are finished building it there isn't going to be anything there but the robot body — the implemented cognitive system. In particular, we are not going to create some kind of mystical non-physical self. Believing this, we may set ourselves the task of enabling the robot to engage in practical and epistemic c-reasoning without having any false c-beliefs to the effect that it is something other than a complicated lump of plastic, silicon, and titanium. What is interesting is — *that cannot be done!* The arguments of section five show that the only way to build a general-purpose robot capable of practical c-reasoning is to provide it with a *de se* c-symbol and program it to use that symbol in certain specific ways in c-reasoning. The result will replicate those aspects of the human cognitive architecture that lead us into the mind/body problem, and if the robot is smart enough they will lead it there as well. That is, the robot will c-conclude that "it" is not a physical robot, and its c-reasoning will be unexceptional when viewed from the perspective of the epistemic norms implemented in its cognitive system. Norms that are necessary to make practical c-reasoning work also lead the robot inexorably to the (false) c-conclusion that there is something there other than the robot.

For example, consider the *cogito* argument with respect to the robot. The robot can run that argument just like we do. It can c-think "I think", and then go on to c-infer that "it" exists and that "it" is not identical with its body, without either of those c-beliefs being true. The *cogito* fails because the robot can have the c-thought "I exist" without there being anything that has the thought "I exist".

The point of this is that the arguments that led us to conclude that we are non-physical selves are not plausible when we apply them in the third-person to the robot. They

just lead us to the conclusion that the robot is wrong in c-believing that "it" (a self distinct from the physical robot) exists. Shouldn't we think that we are like the robot? The same computational pressures that lead the robot to c-believe that it exists lead us to believe that we exist. If there is no reason to think that there is anything there in the robot to make its c-belief true, shouldn't we be equally dubious about ourselves?

7. Is This Intelligible?

The preceding arguments are, we feel, strong. But the conclusions are perverse. It would be irrational for me to conclude "I do not exist". My conceptual framework mandates believing various things about myself, such as my location relative to my visual field. It follows immediately from these beliefs that I exist. E.g., if I am at a certain location then I exist. I cannot, rationally, give up the belief that I exist.

On the other hand, something similar is true of the robot, but we are inclined to say that the robot's c-belief is false. It makes a difference whether we are thinking of rational agents from the inside or the outside. Thinking of our robot from the outside, we can insist that it is wrong in c-believing "I exist", but we can simultaneously insist that it would be irrational for the robot to c-believe otherwise. But when I think about myself, I cannot get outside of my own epistemic norms. Judging that the robot is rational in c-thinking that it exists is tantamount in myself to simply c-thinking that I exist. I cannot, rationally, do otherwise.

This remains perplexing, however. It does not seem reasonable to conclude that the robot has somehow come to embody a non-physical self. In order for that to be the case, it would have to be its computational organization that somehow brings that non-physical self into existence, but how could that be? On the other hand, assuming that *I* do exist, it seems that the only possible explanation for this would be that *I am* brought into existence by my body's computational organization. And if I am willing to say this about myself, why should I be reluctant to say it about the robot?

Our conclusion is that we really don't know what to conclude. We lay these arguments out for your perusal, and you should draw your own conclusions (if you exist).

8. Conclusions and Comparisons

This paper has two parts — a constructive part and a skeptical part. In the constructive part we investigate the logical role of reflexive designators in rational cognition. There

is a rich literature on reflexive designators, going back to Castañeda and Perry. The original interest in reflexive designators was from the perspective of the philosophy of language. Pollock briefly investigated their role in practical cognition in "My brother the machine" (*Nous* **22** (1988), 173-212). Perry takes the issue up in some of his recent work,[10] arriving at similar conclusions to Pollock. We have taken the matter further here, correcting various aspects of earlier proposals and arguing that purely computational pressures deriving from the logical structure of rational cognition dictate the need for reflexive designators in sophisticated agents. Furthermore, we have argued that different aspects of rational cognition give rise to the different (temporal, spatial, and personal) reflexive designators.

The skeptical part of the paper derives from the observation that *de se* designators could serve their requisite functional role in cognition without actually designating anything. If we incorporate such designators into the cognition of a robot, there is reason to be skeptical about their designating anything. But then, why shouldn't we be equally skeptical about our own existence? The difficulty is that we cannot be skeptical about our own existence. Our epistemic norms do not allow that. So we have what seems to be a fairly strong argument for nolipsism, but it is not a view we can endorse.

Dennett is well known for having suggested related views, but upon close inspection it is not clear what his views actually are. In "Where am I?",[11] he proposes a range of cases in which the brain, the body, and the center of the perceptual perspective come apart. He doesn't explicitly draw conclusions, but seems to gravitate towards locating himself where his brain is. The general recipe for generating these kinds of cases is clear enough. Those considered by Dennett involve an embarrassment of riches — too many material candidates for selfhood, i.e., too many space-occupying hunks of matter with a claim to being me, and in the end, a case in which there are a pair of selves, located in the same place. We add to the pot a range of cases in which there are no natural material candidates for selfhood, nothing that occupies the center of my perceptual perspective, no enduring hunk of matter whose actions I control, nor even, necessarily, anything like a brain, a spatially localized bit of organic matter where computation goes on, and which serves as the causal basis for my experience. And, unlike Dennett, we give arguments from the functional role of *de se* designators rather

---

[10] "Self-notions", *Logos* 1990, 17-31, and "Myself and I" in *Philosohie in Synthetisher Asicht*, ed. Macello Stamm, Stuttgart: Klett-Cotta, 1998, 83-103.

[11] In *Mind's I*, ed. Douglas Hofstadter and Daniel Dennett, Basic Books, 1981.

than merely telling stories.

It is unclear what Dennett wants to conclude from his stories. In some places, he describes the self as a kind of fiction; a unified agent posited as part of a hermeneutic activity designed to explain our own behavior. He writes:

> We are all virtuoso novelists, who find ourselves engaged in all sorts of behavior, more or less unified, but sometimes disunified, and we always put the best "faces" on it we can. We try to make all of our material cohere into a single good story. And that story is our autobiography. The chief fictional character at the center of that autobiography is one's self.[12]

But in other places he says that the self is perfectly real, but abstract — an organization that tends to distinguish, control and preserve portions of the world.[13]

Nolipsism is perhaps most closely allied with a class of views associated with Lichtenberg, Wittgenstein, Anscombe, and sometimes Schlick, for which Strawson coined the term *no-subject views*.[14] But it is important to emphasize that we stop short of endorsing nolipsism. The point of the skeptical part of the paper is simply that the earlier constructive account of the functional role of *de se* designators provides the basis for what seems to be a rather strong argument for nolipsism. One would normally be inclined to endorse such a strong argument were it not that our own computational structure (our epistemic norms) make it impossible for us to accept the conclusion.

---

[12] "The self as a center of narrative gravity", in *Self and Consciousness: Multiple Perspectives*, ed. F. Kessel, P. Cole, and D. Johnson, Hillsdale, NJ: Erflbaum, 1992.)

[13] "The origins of selves: Do I choose who I am?", *Cogito* (1989), 163-173.

[14] *Individuals*, London: Metheun, 1959, pg 95.