Preprint

This chapter is forthcoming in the *Journal of Experimental and Theoretical Artificial Intelligence*.

Luciano Floridi

Digitally signed by Luciano Floridi
DN: CN = Luciano Floridi, C = GB, OU = Oxford University
Reason: I am the author of this document
Date: 2005.09.23 00:07:14 +02'00'

Solving the Symbol Grounding Problem: a Critical Review of Fifteen Years of Research

Mariarosaria Taddeo[1] and Luciano Floridi[1,2]

[1] Dipartimento di Scienze Filosofiche, Facoltà di Lettere e Filosofia, Università degli Studi di Bari, Italy. E-mail: mariarosaria.taddeo@filosofia.uniba.it

[2] Faculty of Philosophy and IEG, Computing Laboratory, Oxford University, Great Britain. E-mail: luciano.floridi@philosophy.oxford.ac.uk

Abstract

This article reviews eight proposed strategies for solving the *Symbol Grounding Problem* (SGP), which was given its classic formulation in Harnad (1990). After a concise introduction, we provide an analysis of the requirement that must be satisfied by any hypothesis seeking to solve the SGP, the *zero semantical commitment condition*. We then use it to assess the eight strategies, which are organised into three main approaches: representationalism, semi-representationalism and non-representationalism. The conclusion is that all the strategies are semantically committed and hence that none of them provides a valid solution to the SGP, which remains an open problem.

Keywords: artificial agent, representationalism, semantical commitment, semantics, symbol grounding problem.

1. The Symbol Grounding Problem

Harnad (1990) uses the Chinese Room Argument (Searle 1980) to introduce the SGP.[1] An artificial agent (AA), such as a robot, appears to have no access to the meaning of the symbols it can successfully manipulate syntactically. It is like someone who is expected to learn Chinese as her first language by consulting a Chinese-Chinese dictionary. Both the AA and the non-Chinese speaker are bound to be unsuccessful, since a symbol may be meaningful, but its mere physical shape and syntactic properties normally provide no clue as to its corresponding semantic value, the latter being related to the former in a notoriously, entirely arbitrary way.

Usually, the symbols constituting a symbolic system neither resemble nor are causally linked to their corresponding meanings. They are merely part of a formal, notational convention agreed upon by its users. One may then wonder whether an AA (or indeed a population of them) may ever be able to develop an *autonomous*, semantic capacity to connect its symbols with the environment in which the AA is embedded interactively. This is the SGP. As Harnad phrases it: "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?" (Harnad 1990, p. 335).

2. Fifteen Years of Research

In this paper, we review eight strategies that have been proposed for solving this *Symbol Grounding Problem* (SGP). We begin by analysing the requirement that must be satisfied by any hypothesis seeking to solve the SGP, the *zero semantical commitment condition*. The latter is then used in the rest of the paper to assess the eight strategies. These are organised into three main approaches: representationalism, semi-representationalism and non-representationalism.

The representationalist approach is discussed in section 4. The first strategy (Harnad 1990) is analysed in section 4.1. It provides the basis for two other strategies (Mayo 2003 and Sun 2000), which are analysed in sections 4.2 and 4.3 respectively.

[1] See also Harnad (2003) for a more recent formulation.

Three semi-representationalist strategies (Vogt 2002a, Davidsson 1995, and Rosenstein and Cohen 1998) are the topic of section 5. They attempt to show that the representations required by any representationalist approach to the SGP can be elaborated in terms of processes implementable by behavioural-based robots. They are assessed in sections 5.1, 5.2 and 5.3 respectively.

The non-representationalist approach is discussed in section 6, where the *Physical Grounding Hypothesis* (Brooks 1990 and 1991) is first recalled. There follows a discussion of two communication- and behaviour-based strategies (Billard and Dautenhahn 1999, Varshavskaya 2002) in sections 6.1 and 6.2 respectively.

All approaches seek to ground the symbols through the sensorimotor capacities of the artificial agents involved. The strategies differ in the methods used to elaborate the data obtained from the sensorimotor experiences and in the role (if any) assigned to the elaboration of the data representations in the process of generating the semantics for the symbols. Unfortunately, all strategies turn out to be semantically committed and hence none of them can be said to offer a valid solution to the SGP.

Three caveats are in order before moving to the next section. First, our goal in this review is to assess the wide range of strategies that have been proposed for solving the SGP in the last fifteen years. It is not to compile an exhaustive bibliography on the SGP, nor to reconstruct the history of the extended literature on this topic. Our claim is only that such literature can fruitfully be organised and assessed as shown in this review.

Second, the works reviewed have been selected for their influential role in several lines of research and/or for their representative nature, insofar as each of them provides an enlightening example of the sort of perspective that might be adopted to tackle the SGP. No inference should be drawn on the scientific value of works which have not been included here, especially since we have focused only on strategies explicitly addressing the SGP, disregarding the debates on

- the *Chinese Room Argument* (Searle 1980), reviewed by Cole (2004);
- the *representation grounding problem* (Chalmers 1992), the *concept grounding problem* (Dorffner and Prem 1993) and the *internalist trap* (Sharkey and Jackson 1994), all reviewed by Ziemke (1999); and
- the *symbols anchoring problem*, reviewed by Coradeschi and Saffioti (2003).

It is worth stressing, however, that the conclusion reached in our review – that symbol grounding is a crucial but still unresolved problem – is consistent with the conclusions reached by Cole, Ziemke and Coradeschi and Saffioti.

Third, although we have tried to provide a coherent and unifying frame of conceptual and technical vocabulary throughout the review, some lack of uniformity has been inevitable owing to the variety of methods, intellectual traditions and scientific goals informing the strategies analysed.

3. The Zero Semantical Commitment Condition

The SGP is one of the most important open questions in the philosophy of information (Floridi 2004). It poses a radical and deceptively simple challenge. For the difficulty is not (or at least, not just) merely grounding the symbols *somehow* successfully, as if all we were looking for were the implementation of some sort of internal lookup table or the equivalent of a searchable spreadsheet. The SGP concerns the possibility of specifying *precisely how* an AA can *autonomously* elaborate its own semantics for the symbols that it manipulates and do so from scratch, by interacting with its environment and other AAs. This means that, as Harnad rightly emphasises, the interpretation of the symbols must be *intrinsic* to the symbol system itself, it cannot be *extrinsic*, that is, *parasitic* on the fact that the symbols have meaning for, or are provided by, an interpreter. It follows that

a) no form of *innatism* is allowed; no semantic resources (some *virtus semantica*) should be presupposed as already pre-installed in the AA; and
b) no form of *externalism* is allowed either; no semantic resources should be uploaded from the "outside" by some *deus ex machina* already semantically-proficient.

Of course, points (a)-(b) do not exclude the possibility that

c) the AA should have its own capacities and resources (e.g. computational, syntactical, procedural, perceptual, educational etc., exploited through algorithms, sensors, actuators etc.) to be able to ground its symbols.

These points only exclude the possibility that such resources may be semantical in the first place, if one wishes to appeal to them in order to solve the SGP without begging the question.

Points (a)-(c) clarify the sense in which a valid solution of the SGP must be fully *naturalised*, despite the fact that we are talking about *artificial* agents. They define a

requirement that must be satisfied by any strategy that claims to solve the SGP. We shall label this the *zero semantical commitment condition* (henceforth Z condition). Any approach that breaches the Z condition is semantically committed and fails to solve the SGP.

We shall now review eight strategies proposed for the solution of the SGP in the last fifteen years. The conclusion will be that none of them satisfies the Z condition. This, of course, does not mean that they are theoretically flawed or uninteresting, nor that they cannot work when technically implemented by actual AAs. But it does mean that, conceptually, insofar as they are successful, such strategies either fail to address the SGP or circumvent it, by implicitly presupposing its solution. In either case, the challenge posed by the SGP is still open.

4. The Representationalist Approach

The representationalist approach considers the conceptual and categorical representations, elaborated by an AA, as the meanings of the symbols used by that AA. So, representationalist strategies seek to solve the SGP by grounding an AA's symbols in the representations arising from the AA's manipulations of its perceptual data. More specifically, it is usually argued that an AA is (or at least should be) able to

1. capture (at least some) salient features shared by sets of perceptual data;
2. abstract them from the data sets;
3. identify the abstractions as the contents of categorical and conceptual representations; and then
4. use these representations to ground its symbols.

The main problem with the representationalist approach is that the available representations – whether categorical or perceptual – succeed in grounding the symbols used by an AA only at the price of begging the question. Their elaboration and hence availability presupposes precisely those semantic capacities or resources that the approach is trying to show to be evolvable by an AA in the first place.

4.1. A Hybrid Model for the Solution of the SGP

Harnad (1990) suggests a strategy based on a *hybrid model* that implements a mixture of features characteristic of symbolic and of connectionist systems.

According to Harnad, the symbols manipulated by an AA can be grounded by connecting them to the perceptual data they denote. The connection is established by a bottom-up, invariantly categorizing processing of sensorimotor signals. Assuming a general psychological theory that sees the ability to build categories[2] of the world as the groundwork for language and cognition (Harnad 1987), Harnad proposes that symbols could be grounded in three stages:

1. *iconization*: the process of transforming analogue signals (patterns of sensory data perceived in relation to a specific entity) into iconic representations (that is, internal analog equivalents of the projections of distal objects on the agent's sensory surfaces);
2. *discrimination*: the process of judging whether two inputs are the same or, if they are different, how much they differ;
3. *identification*: the process of assigning a unique response – that is, a name – to a class of inputs, treating them as equivalent or invariant in some respect.

The first two stages yield sub-symbolic representations; the third stage grounds the symbols. The iconic representations in (1) are obtained from the set of all the experiences related to the perceptions of the *same type* of object. The categorical representations are then achieved through the discrimination process in (2). Here, an AA considers only the *invariant features* of the iconic representations. Once elaborated, the categorical representations are associated in (3) with classes of symbols (the names), thus providing the latter with appropriate referents that ground them.

Iconization and *discrimination* are sub-processes, carried out by using neural networks. They make possible the subsequent association of a name with a class of input and subsequently the naming of referents. However, by themselves neural networks are unable to produce symbolic representations, so they cannot yet enable the AA to develop symbolic capacities. In order to avoid this shortcoming, Harnad provides his hybrid model with a symbolic system, which can manipulate symbols syntactically and finally achieve a semantic grounding of its symbols.

Harnad's proposal has set the standard for all following strategies. It attempts to overcome the typical limits encountered by symbolic and connectionist systems by

[2] Harnad uses the term *category* to refer to the name of the entity denoted by symbol, so a category is not itself a symbol. A grounded symbol would have both categorical (i.e. a name) and iconic representations.

combining their strengths. On the one hand, in "a pure symbolic model the crucial connection between the symbols and their referents is missing; an autonomous symbol system, though amenable to a systematic syntactic interpretation, is ungrounded" (Harnad 1990, p. 341-342). On the other hand, although neural networks make it possible to connect symbols and referents by using the perceptual data and the invariant features of the categorical representations, they still cannot manipulate symbols (as the symbol systems can easily do) in order to produce an intrinsic, systematic and finite interpretation of them; hence the hybrid solution supported by Harnad, which, owing to its semi-symbolic nature, may seem to represent the best of both worlds.

Unfortunately, the hybrid model does not satisfy the Z condition. The problem concerns the way in which the hybrid system is supposed to find the invariant features of its sensory projections that allow it to categorize and identify objects correctly. Consider an AA that implements the hybrid model, called PERC ("PERCeives"). Initially, PERC has no semantic contents or resources, so it has no semantical commitment. PERC is equipped with a digital video camera, through which it observes its external environment. Following Harnad, suppose that, by means of its camera and neural networks, PERC is able to produce iconic representations from the perceptual data it collects from the environment. PERC is then supposed to develop categorical representations from these perceptual data, by considering only the invariant features of the iconic representations. Next, it is supposed to organize the categorical representations into conceptual categories, like "quadruped animal". The latter are the meanings of the symbols. The question to be asked is where conceptual categories such as "quadruped animal" come from. Neural networks can be used to find structures (if they exist) in the data space, such as patterns of data points. However, if they are *supervised*, e.g. through *back propagation*, they are trained by means of a pre-selected training set and repeated feedback, so whatever grounding they can provide is entirely extrinsic. If they are *unsupervised*, then the networks implement training algorithms that do not use desired output data but rely only on input data to try to find structures in the data input space. Units in the same layer compete with each other to be activated. However, they still need to have *built-in biases* and *feature-detectors* in order to reach the desired output. Such semantic resources are necessarily hard-coded by a supervisor, according to pre-established criteria. Moreover, unsupervised or self-organizing networks, once they have been trained,

still need to have their output checked to see whether the obtained structure makes any sense with respect to the input data space. This difficult process of validation is carried out externally by a supervisor. So in this case too, whatever grounding they can provide is still entirely extrinsic. In short, as Christiansen and Chater (1992, p. 235) correctly remark "[So,] whatever semantic content we might want to ascribe to a particular network, it will always be parasitic on our interpretation of that network; that is, parasitic on the meanings in the head of the observer".

"Quadruped animal", as a category, is not the outcome of PERC's intrinsic grounding because PERC must already have had quite a lot of semantic help to reach that conclusion. The strategy supported by Harnad actually presupposes the availability of those semantic resources that the AA is expected to develop from scratch, through its interactions with the environment and the elaboration of its perceptual data.

It might be retorted that the categorical representations do not need to collect all the invariant features of the perceptual data, for they may just indicate a class of similar data, which could then be labelled with a conventional name. Allegedly, this could allow one to avoid any reliance on semantical resources operating at the level of the neural network component. The reply resembles Berkeley's criticism of Locke's semantic theory of *general* or *abstract ideas*.

Locke had suggested that language consists of conventional signs, which stand for simple or abstract ideas. Abstract ideas, such as that of a horse, correspond to general names, e.g. "horse", and are obtained through a process of abstraction not dissimilar from the process that leads to categorical representations in Harnad's hybrid model, that is, by collecting the invariant features of simple ideas, in our case the many, different horses perceivable in the environment.

Against Locke's theory, Berkeley objected that the human mind elaborates only particular ideas (ideas of individuals, e.g. of that specific white and tall and ... horse, or this peculiar brown, and short and... horse, and so forth) and therefore that universal ideas and the corresponding general names, as described by Locke, were impossible. This is especially true for abstract universal ideas. For example, the idea of "extension", Berkeley argued, is always the idea of something that is extended. According to Berkeley, universal or abstract ideas are therefore only particular ideas that (are chosen to) work like prototypes or models standing for a class of similar but

equally particular ideas. In this way, the idea of a specimen is elected to the role of abstract idea of the whole class to which the specimen belongs.

Returning to Harnad, although he suggests that the categories available to an AA are the consequence of a Lockean-like abstraction from perceptual data, one may try to avoid the charge of circularity (recall that the solution has been criticised for infringing the Z condition) by trying to redefine the categorical representation in more Berkeleian terms: a particular representation could be used by an AA as a token in order to represent its type.

Unfortunately, this Berkeleian manoeuvre does not succeed either. For even if categorical representations – comparable to Lockean abstract ideas – are reduced to iconic representations – comparable to Berkeleian abstract ideas – the latter still need to presuppose some semantic resources to be elaborated. In our example, how is the class of horses (the data space) put together in the first place, without any semantic capacity to elaborate the general idea (whether Lockean or Berkeleian does not matter) of "horse" to begin with? And how is a particular specimen of horse privileged over all the others as being *the* particular horse that could represent all the others? And finally, how does one know that what makes that representation of a particular horse the representation of a universal horse is not e.g. the whiteness instead of the four-legged nature of the represented horse? The Z condition is still unsatisfied.

In sections 4.2 and 4.3, we shall assess two other solutions of the SGP based on Harnad's. Both raise some new difficulties. Before that, however, we shall briefly look at the application of Harnad's solution to the explanation of the origin of language and its evolution, in section 4.1.1. The topic has been investigated by Harnad himself on several occasions. Given the scope of this review, we shall limit our discussion to three papers: Cangelosi, et al. (2000), Harnad and Cangelosi (2001) and Cangelosi, et al. (2002). These are based on Harnad (1990). They maintain that, within a plausible cognitive model of the origin of symbols, symbolic activity should be conceived as some higher-level process, which takes its contents from some non-symbolic representations obtained at a lower level. This is arguably a reasonable assumption. Because of their reliance on Harnad's initial solution, however, the papers share its shortcomings and are subject to the same criticism. They are all semantically committed and hence none of them provides a valid solution for the SGP.

The three papers show that, despite Harnad's (1993) reply to Christiansen and Chater (1992), in subsequent research Harnad himself has chosen to follow a non-deflationist interpretation of his own solution of the SGP.[3] However, it seems that either Harnad's reply to the objection moved by Christiansen and Chater is satisfactory, but then Harnad's strategy for solving the SGP becomes too general to be of much interest; or Harnad's strategy is a substantial, semantic proposal, in which case it is interesting but its is also subject to the objection in full.[4]

4.1.1. SGP and the Symbolic Theft Hypothesis

Cangelosi and Harnad (2001) and Cangelosi, et al. (2000) provide a detailed description of the mechanisms for the transformation of *categorical perception* (CP) into grounded, low-level labels and, subsequently, into higher-level symbols.[5] They call *grounding transfer* the phenomenon of acquisition of new symbols from the combination of already-grounded symbols. And they show how such processes can be implemented with neural networks: "Neural networks can readily discriminate between sets of stimuli, extract similarities, and categorize. More importantly, networks exhibit the basic CP effect, whereby members of the same category "look" more similar (there is a compression of within-category distances) and members of different categories look more different (expansion of between-categories distances)." (Cangelosi, et al. 2002, p. 196).

According to Cangelosi and Harnad (2001), the functional role of CP in symbol grounding is to define the interaction between *discrimination* and *identification*. We have seen in 4.1 that the process of discrimination allows the system to distinguish patterns in the data, whilst the process of identification allows it to assign a stable identity to the discriminated patterns. "CP is a basic mechanism for providing more compact representations, compared with the raw sensory projections where feature-filtering has already done some of the work in the service of categorization." (Cangelosi, et al. 2002, p. 198).

---

[3] A deflationist view of the SGP is supported by Prem (1995a,b,c), who argues that none of the different approaches to the problem of grounding symbols in perception succeed in reaching its semantic goals and that SG systems should rather be interpreted as some kind of automated mechanisms for the construction of models, in which the AA uses symbols to formulate descriptive rules about what will happen in its environment

[4] Although for different reasons, a similar conclusion is reached by Taylor and Burgess (2004).

[5] [5] The same mechanism is also described in Cangelosi (2001) and Harnad (2002).

Cangelosi, et al. (2000) outline two methods to acquire new categories. They call the first method *sensorimotor toil* and the second one *symbolic theft*, in order to stress the benefit (enjoyed by the system) of not being forced to learn from a direct sensorimotor experience whenever a new category is in question.

Cangelosi, et al. (2000) provide a simulation of the process of CP, of the acquisition of grounded names, and of the learning of new high-order symbols from grounded ones. Their simulation comprises a three-layer feedforward neural network, which has two groups of input units: forty-nine units simulating a retina and six units simulating a linguistic input. The network has five hidden units and two groups of output units replicating the organization of input (retina and verbal output). The retinal input depicts nine geometric images (circles, ellipses, squares, rectangles) with different sizes and positions. The activation of each input unit corresponds to the presentation of a particular category name. The training procedure (which is problematic in view of the Z condition) has the following learning stages:

1) the network is *trained by an external agent already semantically proficient*; (so this breaches the Z condition) to categorize figures: from input shapes it must produce the *correct* (here hides another breach of the Z condition) categorical prototype as output;
2) the network is then given the task of associating each shape with its name. This task is called *entry-level naming*. According to the authors, names acquired in this way can be considered grounded because they are explicitly connected with sensory retinal inputs. However, the semantic commitment is obvious in the externally supervised learning process;
3) in the final stage, the network learns how to combine such grounded names (for example, “square” or “rectangle”) with new arbitrary names (for example “symmetric” or “asymmetric”). This higher-level learning process is implemented by simple imitation learning of the combination of names. This is like teaching the system conceptual combinations such as “square is symmetric” or “rectangle is asymmetric”. The AA learns through the association of grounded names with new names, while the grounding is *transferred* to names that did not have such a property.

The model has been extended to use the combination of grounded names of basic features in order to allow systems to learn higher-order concepts. As the authors comment “[T]he benefits of the symbolic theft strategy must have given these

organisms the adaptive advantage in natural language abilities. This is infinitely superior to its purely sensorimotor precursors, but still grounded in and dependent on them" (Cangelosi, et al. 2002, p. 203).

The explanation of the origin and evolution of language, conjectured by this general approach, is based on the hybrid symbolic/sensorimotor capacities implemented by the system. Initially, organisms evolve an ability to build some categories of the world through direct sensorimotor toil. They also learn to name such categories. Then some organisms must have experimented with the propositional combination of the names of these categories and discover the advantage of this new way of learning categories, thus "stealing their knowledge by hearsay" (Cangelosi, et al. 2002, p. 203). However, the crucial issue of how organisms might have initially learnt to semanticise the data resulting from their sensorimotor activities remains unsolved, and hence so does the SGP.

4.2. A Functional Model for the Solution of the SGP

Mayo (2003) suggests a *functional model* of AA that manages to overcome some of the limits of Harnad's hybrid model, although it finally incurs equally insurmountable difficulties.

Mayo may be interpreted as addressing the objection, faced by Harnad (1990), that an AA fails to elaborate its semantic categories autonomously. His goal is to show that an AA could elaborate concepts in such a way as to be able to ground even abstract names.

An AA interacting with the environment perceives a continuum of sensory data. However, data always underdetermine their structure, so there is a countless variety of possible categories (including categories related to particular tasks) by means of which the data could be organized. As Mayo acknowledges "[...] without some sort of bias, it is computationally intractable to come up with the best set of categories describing the world. [...] given that sensory data is continuous, there is an effectively infinite [...] number of possible categorizations of the data." (Mayo 2003, p. 56). So Mayo proposes a *functional organization* of the representations as a way to ground the symbols involved. Categories are interpreted as *task-specific sets* that collect representations according to their practical function. Symbols are formed in order to solve specific task-oriented problems in particular environments. Having a specific task to perform provides the AA with a bias that orientates its search for the best

categorisation of sensory data. The bias is such that the symbols learnt by the AA are those that most help the AA to perform the task successfully. A symbol could then acquire different meanings, depending on the functional set in which it occurs. The sets overlap insofar as they share the same symbols and, according to Mayo, these intersections support the capacity of the AA to generalize and to name abstract concepts. For example, an AA can generalize the meaning of the symbol 'victory' if, according to Mayo, 'victory' is not rigidly connected to a specific occurrence of a single event but derives its meaning from the representation of the intersection of all the occurrences of "victory" in different task-specific sets of various events, such as "victory" in chess, in tennis, in war and in love.

Contrary to the hybrid model, the functional model avoids the problem concerning the elaboration of abstract concepts by the AA. However, like all the other representationalist hypotheses, Mayo's too founds the elaboration of the semantics on categorical and symbolic representations. But then, as in Harnad (1990), the initial presence of these representations requires the presence of substantial semantic capacities that cannot simply be warranted without begging the question. In Mayo's case, these are the functional criteria. The AA is already presumed to have (access to, or the capacity to generate and handle) a "functional" semantics. The AA is not (indeed it cannot be) supposed or even expected to elaborate this semantic resource by itself. Obviously, the strategy is already semantically committed and such commitment undermines its validity.

The difficulty might be avoidable by a model in which some *internal* (or internally developed) semantic resource allows the AA to organize its categories functionally and hence to ground its symbols *autonomously*. A proposal along these lines has been developed by Sun (2000), as we shall see in the next section.

4.3. An Intentional Model for the Solution of the SGP

Sun (2000) proposes an *intentional model* that relates connectionism, symbolic representations and situated artificial intelligence.[6] As for Harnad and Mayo, for Sun too the AA's direct interaction with the environment is pivotal in the elaboration of its symbolic representations and hence the solution of the SGP. The novelty lies in the development by the AA of some intentional capacities.

[6] The strategy is developed in several papers, see Sun (1997), Wermter and Sun (2000), Sun (2001a), Sun (2001b), Sun, et al. (2001), Slusarz and Sun (2001), Sun and Zhang (2002).

Sun refers to the interaction between an AA and the environment in the Heideggerian terms of *being-in-the-world* and *being-with-the-world*. As he remarks, "[the ability to elaborate] the representations presupposes the more basic comportment [of the agent] with-the-world." (Sun 2000, p. 164). The AA is *in-the-world* and interacts with objects in the world in order to achieve its goals. Its intentional stance is defined in the still Heideggerian terms of *being-with-the-things*.

According to Sun, representations do not stand for the corresponding perceived objects but for the uses that an AA can make of these objects as means to ends. The intentional representations contain the rules for the teleological use of the objects and the AA elaborates this kind of representations through a learning process.

Still following a Heideggerian approach, Sun distinguishes between a first and a second level of learning: "it is assumed that the cognitive processes are carried out in two distinct levels with qualitatively different processing mechanisms. Each level encodes a fairly complete set of knowledge for its processing." (Sun 2002, p. 158)

The two levels complement each other. The first-level learning directly guides the AA's actions in the environment. It allows the AA to follow some courses of action, even if it does not yet know any rule for achieving its goals. At this stage, the AA does not yet elaborate any explicit representations of its actions and perceptual data. The first-level learning guides the behaviour of the AA by considering only two factors: the structure of the external world and the "*innate biases* or *built-in constraints* and *predispositions* [emphasis added] which also depend on the (ontogenetic and phylogenetic) history of agent world interaction." (Sun 2000, p. 158). Such an innate criterion – which already breaches the Z condition – is identified by Sun with a first-level intentionality of the AA, which is then further qualified as "pre-representational (i.e., *implicit*)" (Sun 2000, p. 157, emphasis added). Such intentionality provides the foundation for the initial interactions of the AA with its environment and for the subsequent, more complex form of intentionality.

During the first-level learning stage, the AA proceeds by trial and error, in order to discover the range of actions that best enable it to achieve its goals. These first-level learning processes allow the AA to acquire the initial data that can then work as input for its second-level learning processes. The latter produce the best possible behaviour, according to some of the AA's parameters, to achieve its objectives. It is at this second-level stage of learning that the AA elaborates its conceptual representations from its first-level data, thanks to what Sun (2000) defines as second-

level intentionality. At the first-level, the behaviour of the AA is intentional in the sense that it directs the AA to the objects in the world. Second-level intentionality uses first-level intentionality data in order to evaluate the adequacy of different courses of action available to the AA to achieve its objectives. According to Sun (2002), this is sufficient to ground the conceptual representations in the AA's everyday activities, in a functional way.

So far, we have described first and second-level learning processes as layered in a bottom-up, dynamic structure but, according to Sun, there is also a top-down dynamic relation among the layers. This allows the AA to generalize the representations obtained in relation to its best behaviours, in order to use them in as many cases as possible. Through a top-down procedure, the AA verifies once more the validity of the representations elaborated, compares the selected representations with the goals to be achieved, generalizes those representations already related to the best behaviours (given some parameters) and fine-tunes the remaining representations to ensure that they are related to a more successful behaviour.

The intentional model elaborated by Sun defines a specific architecture for the AA, which has been implemented in a system called CLARION (Sun and Peterson 1998). We shall briefly describe its features in order to clarify the difficulties undermining Sun's strategy for solving the SGP.

4.3.1. CLARION

CLARION consists of four layered neural networks (see the problem in using neural networks to solve the SGP, discussed in section 4.1), which implement a bottom-up process. The first three levels elaborate the values of CLARION's actions. The fourth level compares the values of the actions and – given some parameters – chooses the best course to achieve its goals, elaborates an explicit rule and adds it to the symbolic level.

To evaluate its actions, CLARION employs a Machine Learning algorithm known as *Q-learning*. This is based on the reinforcement learning principle. Suppose an AA is confronted by a specific task. The algorithm models the task in terms of states of the AA and actions that the AA can implement starting from its current state. Not all states lead to the goal state, and the agent must choose a sequence of optimal or sub-optimal actions that will lead to the goal state, by using the least possible states to minimize cost. Each good choice is rewarded and each bad choice is punished. The

agent is left training on its own, following these rules and rewards. During the training process, the agent learns what the best actions are to achieve a specific task. Given sufficient training time, the agent can learn to solve the problem efficiently. Note, however, that the algorithm works only if the problem can be modelled and executed by an algorithm in a finite time because the number of states and actions are relatively finite. A game like Go is already too complex. As far as the solution of the SGP is concerned, it is already clear that, by adopting the Q-learning algorithm, the intentional model is importing from the outside the very condition that allows CLARION to semanticise, since tasks, goals, success, failure, rewards and punishments are all established by the programmer. The semantical commitment could not be more explicit.

CLARION's symbolic fourth level corresponds to the second-level learning process in Sun's model. The values of the actions are checked and generalized in order to make possible their application even in new circumstances. This last stage corresponds to the top-down process. CLARION's high-level concepts are "context dependent and they are functional to achieve the objectives of the agents […] the concepts are part of the set of roles which an agent learns in order to interact with the environment", (Sun 2000, p. 168). Sun stresses the functional nature of the concepts in order to point out that they come from experience and are not defined a priori.

The functionalism implemented by the intentional model is possible only thanks to extrinsic, semantic resources, freely provided to the AA. This undermines the value of Sun's strategy as a solution of the SGP. Sun (2000) attempts to overcome this difficulty by reinterpreting the functionalist criterion as an innate and intrinsic feature of the AA, namely its intentionality. Yet, this alleged solution equally begs the question, since it remains unclear how the AA is supposed to acquire the necessary intentionality without which it would be unable to ground its data. In this case too, semantics is made possible only by some other semantics, whose presence remains problematic.

It might be replied that the intentionality of the representations can arise from the process of extraction of conceptual representations from first-level learning processes and that, at this level, the AA's intentionality could derive from its direct interactions with the world, encoded through its first-level learning. In this way, the semantic resources, to which the AA freely and generously helps itself, would not have to be extrinsically generated. Indeed, Sun (2000) describes first-level intentionality as a

pure consequence of the interactions of an AA with its environment. “Comportment carries with it a direct and an unmediated relation to things in the world […]. Therefore it provides [an] *intrinsic* intentionality (meanings), or in other words a connection (to things with the words) that is intrinsic to the agent […]” (Sun 2000, p. 164). Unfortunately, it remains unexplained precisely how this first-level intentionality might arise in the first place. Presupposing its presence is not an answer. How does even a very primitive, simple and initial form of intentionality develop (in an autonomous way) from the direct interactions between and AA and its environment? Unless a logically valid and empirically plausible answer is provided, the SGP has simply been shifted.

Sun (2000) argues that AAs evolve, and hence that they may develop their intentional capacities though time. In this way, first-level intentionality and then further semantic capacities would arise from evolutionary processes related to the experience of the AAs, without the presence of extrinsic criteria. “There are some existing computational methods available to accomplish simple forms of such [i.e. both first- and second-level] learning. […] [A]nother approach, the genetic algorithm [...] may also be used to tackle this kind of task.” (Sun 2000, p. 160). However, in this case too, the solution of the SGP is only shifted. The specific techniques of artificial evolution to which Sun refers (especially Holland 1975) do not grant the conclusion that Sun’s strategy satisfies the Z condition. Quite the opposite. Given a population of individuals that evolve generationally, evolution algorithms make it possible to go from an original population of “genotypes” to a new generation using only some kind of artificial selection. Evolution algorithms are obviously based on a Darwinian survival mechanism of the fittest. But it is the programmer who plays the key role of the “natural” selection process. She chooses different kinds of “genotype” – AAs with different features – situates them in an environment, calculates (or allows the system to calculate) which is the behaviour that best guarantees survival in the chosen environment, and does so by using a parameter, defined by a *fitness formula*, that once again is modelled and chosen by her. The AAs showing the best behaviour pass the selection, yet “artificial evolutionism” is only an automatic selection technique based on a programmer’s criteria. True, it may possible to hypothesize a generation of AAs that ends up being endowed with the sort of intentionality required by Sun’s strategy. By using the right fitness formula, perhaps a programmer might ensure that precisely the characteristics that allow the AAs to behave in an “intentional way” will

be promoted by their interactions with the environment. For example, a programmer could try to use a fitness formula such that, in the long run, it privileges only those AAs that implement algorithms like CLARION's *Q-learning algorithm*, thus generating a population of "intentional" AAs. However, their intentionality would not follow from their *being-in-the-world*, nor would it be developed by the AAs evolutionary and autonomously. It would merely be superimposed by the programmer's purposeful choice of an environment and of the corresponding fitness formula, until the AAs obtained satisfy the sort of description required by the model. One may still argue that the semantics of the AAs would then be grounded in their first-level intentionality, but the SGP would still be an open challenge. For the point, let us recall, is not that it is impossible to engineer an AA that has its symbols semantically grounded somehow. The point is how an AA can ground its symbols autonomously.

Artificial evolutionism, at least as presented by Sun, does not allow us to consider intentionality an autonomous capacity of the AAs. On the contrary, it works only insofar as it presumes the presence of a semantical framework, from the programmer acting as a *deus ex machina* to the right fitness formula. Sun's strategy is semantically committed and does not provide a valid solution for the SGP.

With the analysis of CLARION we conclude the part of this paper dedicated to the representationalist approach to the SGP. None of the strategies discussed so far appears to provide a valid solution for the SGP. Perhaps the crucial difficulty lies in the assumption that the solution must be entirely representationalist. In the following section we are going to see whether a weakening of the representationalist requirement may deliver a solution of the SGP.

5. The Semi-representationalist Approach

In this section, we review three strategies developed by Davidsson (1995), Vogt (2002a) and Rosenstein and Cohen (1998a). They are still representationalist in nature but differ from the ones discussed in the previous section in that they deal with the AA's use of its representations by relying on principles imported from behaviour-based robotics.

5.1. An Epistemological Model for the Solution of the SGP

According to Davidsson (1995), there is a question that the solution of the SGP suggested by Harnad (1990) leaves unanswered, namely what sort of learning neural

networks allow. We have seen that this issue is already raised by Christiansen and Chater (1992).

Davidsson argues that concepts must be acquired in a gradual fashion, through repeated interactions with the environment over time. The AA must be capable of *incremental* learning, in order to categorize its data into concepts. However, neural networks provide a discriminative learning framework that does not lent itself to an easily incremental adaptation of its contents, given the "fixed-structure of the neural nets" (Davidsson 1995, p. 160). It follows that, according to Davidsson, most neural networks are not suitable for the kind of learning required by an AA that might successfully cope with the SGP. Davidsson (1995) maintains that the SGP becomes more tractable if it is approached in terms of general "conceptual representations" and Machine Learning.

According to Davidsson, "a concept is represented by a composite description consisting of several components." (Davidsson 1995, p. 158). The main idea is that a concept must be a complete description of its referent object, and thus it should collect different kinds of representations, one for each purpose for which the object represented can be used. Davidsson defines three parts of a description:

1. the *designator*, which is the name (symbol) used to refer to a category;
2. the *epistemological representation*, which is used to recognize instances of a category; and
3. the *inferential representation*, which is a collection of all that it is known about a category and its members ("encyclopedic knowledge") and that can be used to make predictions or to infer non-perceptual information.

For example, the concept corresponding to the word "window" could denote a 3-D object model of a typical window and work as an epistemological representation. By means of the inferential knowledge component, one could then include information like: windows are made of wood and glass, they are used to admit light and air in a building, they are fitted with casements or sashes containing transparent material (e.g. glass) and capable of being opened and shut, and so forth.

The epistemological representations are pivotal in Davidsson's solution. They are elaborated through a vision system that allows the identification (categorization) of the perceived data. When an AA encounters an object, it matches the object with its epistemological representation. In so doing, the AA activates a larger knowledge structure, which allows it to develop further, more composite concepts. An

epistemological representation does not have to be (elaborated through) a connectionist network, since it can be any representation that can be successfully used by the vision system to identify (categorize) objects.

Davidsson acknowledges that the representations that ground the symbols should not be pre-programmed but rather learned by the AA from its own "experience". So he suggests using two paradigms typical of Machine Learning: *learning by observation* and *learning from examples*.

Learning by observation is an unsupervised learning mechanism, which allows the system to generate descriptions of categories. Examples are not pre-classified and the learner has to form the categories autonomously. However, the programmer still provides the system with a specific number of well-selected description entities, which allow the AA to group the entities into categories. Clearly, the significant descriptions first selected and then provided by the human trainer to the artificial learner are an essential condition for any further categorization of the entities handled by the AA. They are also a *conditio sine qua non* for the solution of the SGP. Since such descriptions are provided *before* the AA develops its semantics capacities and *before* it starts to elaborate any sort of description autonomously, they are entirely external to the AA and represent a semantical resource implanted in the AA by the programmer.

The same objection applies to the learning from examples mechanism. Indeed, in this case the presence of external criteria is even more obvious, since the sort of learning in question presupposes a set of explicitly pre-classified (by the human teacher) examples of the categories to be acquired. The result is that Davidsson's strategy is as semantically committed as all the others already discussed, so it too falls short of providing a valid solution of the SGP.

5.2. The Physical Symbol Grounding Problem

Vogt (2002a) and Vogt (2002b) connect the solution proposed by Harnad (1990) with situated robotics (Brooks 1990 and 1991) and with the semiotic definition of symbols (Peirce 1931-1958). His strategy consists in approaching the SGP from the vantage point of embodied cognitive science: he seeks to ground the symbolic system of the AA in its sensorimotor activities, transform the SGP into the *Physical Symbol Grounding Problem* (PhSGP), and then solve the PhSGP by relying on two conceptual tools: the *semiotic symbol systems* and the *guess game*.

Vogt defines the symbols used by an AA as a structural pair of sensorimotor activities and environmental data. According to a semiotic definition, AA's symbols have (see Figure 1)

1. a *form* (Peirce's "representamen"), which is the physical shape taken by the actual sign;
2. a *meaning* (Peirce's "interpretant"), which is the semantic content of the sign; and
3. a *referent* (Peirce's "object"), which is the object to which the sign refers.

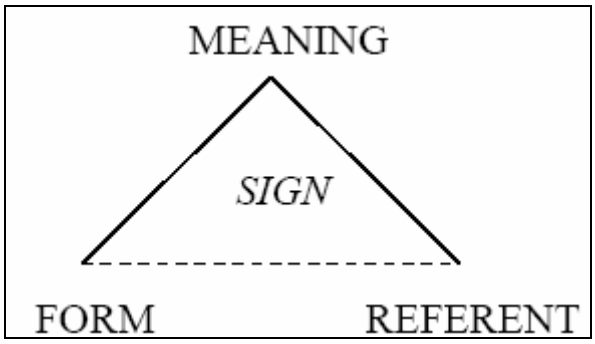

Figure 1 "The semiotic triangle illustrates the relations that constitute a sign. When the form is either arbitrary or conventionalized, the sign can be interpreted as a symbol." (Vogt (2002a, p. 433).

Following this Peircean definition, a symbol always comprises a form, a meaning and a referent, with the meaning arising from a functional relation between the form and the referent, through the process of semiosis or interpretation. Using this definition, Vogt intends to show that the symbols, constituting the AA's semiotic symbol system, are already semantically grounded because of their intrinsic nature. Since both the meaning and the referent are already embedded in (the definition of) a symbol, the latter turns out (a) to be directly related to the object to which it refers and (b) to carry the corresponding categorical representation. The grounding of the whole semiotic symbol system is then left to special kinds of AA that are able to ground the meaning of their symbols in their sensorimotor activities, thus solving the PhSGP.

The solution of the PhSGP is based on the *guess game* (Steels and Vogt 1997), a technique used to study the development of a common language by situated robots.

The guess game (see Figure 2) involves two robots, situated in a common environment. Each robot has a role: the *speaker* names the objects it perceives, the

hearer has the task of finding the objects named by the speaker by trials and error. During the game, the robots develop a common system of semiotic symbol through communicative interactions, the *adaptative language games*. The robots have a very simple body and can only interact with their environment visually. The *speaker* communicates only to convey the name of a visually detected referent. The *hearer* communicates only to inform the speaker about its guessing concerning the referent named by the speaker. The guess game ends successfully if the two robots develop a shared lexicon, grounded in the interactions among themselves and with their environment.

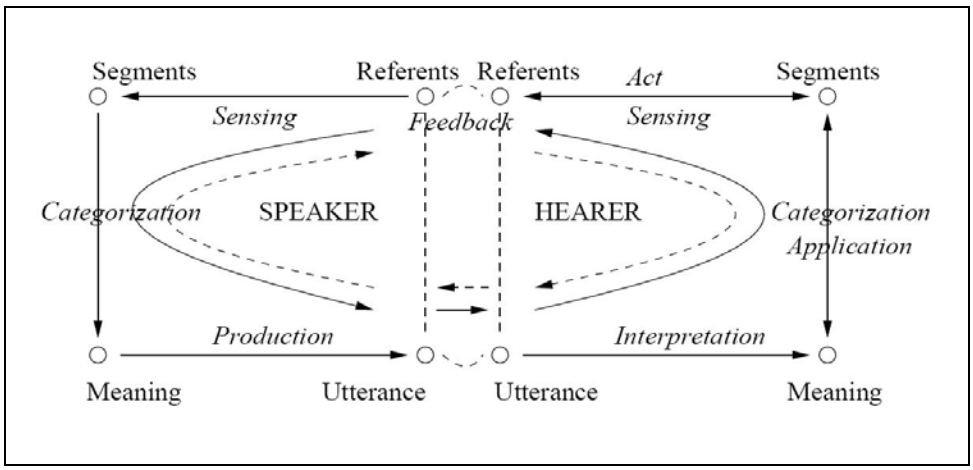

Figure 2 The guess game. "The semiotic square illustrates the guessing game scenario. The two squares show the processes of the two participating robots. This figure is adapted from (Steels and Kaplan 1999)." (Vogt 2002a, p. 438).

The game has four stages, at the end of which the robots are expected to obtain a shared name for an object in their environments.

The first two stages – the beginning of the perceptual activities by the two robots in the environment and the selection of one part of the environment on which they will operate – lie outside the scope of this paper so they will not be analyzed here (for a complete description see Vogt 2002a).

The last two stages concern the processes of meaning formation. More specifically, they constitute the *discrimination game*, through which the categories are elaborated, and the *naming game*, through which the categories are named. These two stages allow the robots to find a referent for their symbols and are crucial for the solution of the SGP.

In order to ground their symbols, the AAs involved in the guess game have to categorize the data obtained from their perception of an object, so that they can later distinguish this category of objects from all the others. According to Vogt, the process for the formation of meaning is carried out by the discrimination game. During this third stage, the AAs – as in Harnad's hybrid model – associate similar perceptual data in order to elaborate their categorical representations. Once the AAs have elaborated one category for each of the objects perceived, the naming game begins. During this last stage, the AAs communicate in order to indicate the objects that they have categorized. The *speaker* makes an utterance that works as the name of one of the categories that it has elaborated. The *hearer* tries to interpret the utterance and to associate it with one of the categories that it has elaborated on its own. The goal is to identify the same category named by the *speaker*. If the *hearer* finds the right interpretation for the *speaker*'s utterance, the two AAs are able to communicate and the guess game is successful.

According to Vogt the guess game makes explicit the meanings of the symbols and allows them to be grounded through the AAs' perceptions and interactions. If the guess game ends successfully, the PhSGP is solved. There are two main difficulties with Vogt's strategy. The most important concerns his semiotic approach; the other relates to what the guess game actually proves.

Suppose we have a set of finite strings of signs – e.g. 0s and 1s – elaborated by an AA. The strings may satisfy the semiotic definition – they may have a form, a meaning and a referent – only if they are *interpreted* by an AA that already has a semantics for that vocabulary. This was also Peirce's view. Signs are meaningful symbols only in the eyes of the interpreter. But the AA cannot be assumed to qualify as an interpreter without begging the question. Given that the semiotic definition of symbols is already semantically committed, it cannot provide a strategy for the solution of the SGP. Nor can the SGP be reduced to the PhSGP: the AA does not have an intrinsic semantics, autonomously elaborated, so one cannot yet make the next move of anchoring in the environment the semantics of the semiotic symbols because there is nothing to anchor in the first place.

It might be replied – and we come in this way to the second difficulty – that perhaps Vogt's strategy could still solve the SGP thanks to the guess game, which could connect the symbols with their external referents through the interaction of the robots with their environment. Unfortunately, as Vogt himself acknowledges, the

guess game cannot and indeed it is not meant to ground the symbols. The guess game assumes that the AAs manipulate previously grounded symbols, in order to show how two AAs can come to make explicit and share the same grounded vocabulary by means of an iterated process of communication. Using Harnad's example, multiplying the number of people who need to learn Chinese as their first language by using only a Chinese-Chinese dictionary does not make things any better.

Vogt acknowledges these difficulties, but his two answers are problematic, and show how his strategy cannot solve the SGP without begging the question. On the one hand, he argues that the grounding process proposed is comparable to the way infants seem to construct meaning from their visual interactions with objects in their environment. However, even if the latter is uncontroversial (which is not), in solving the SGP one cannot merely assume that the AA in question has the semantic capacities of a human agent. To repeat the point, the issue is how the AA evolves such capacities. As Vogt (2002a) puts it, several critics have pointed out that "robots cannot use semiotic symbols meaningfully, since they are not rooted in the robot, as the robots are designed rather than shaped through evolution and physical growth [...], whatever task they [the symbols used by the robots] might have stems from its designer or is in the head of a human observer" (p. 234). To this Vogt replies (and we come in this way to his second answer) that "it will be *assumed* [emphasis added] that robots, once they can construct semiotic symbols, do so meaningfully. This assumption is made to illustrate how robots can construct semiotic symbols meaningfully" (p. 234). The assumption might be useful in order to engineer AAs, but it certainly begs the question when it comes to providing a strategy for solving the SGP.[7]

5.3. A Model based on Temporal Delays and Predictive Semantics for the Solution of the SGP

As in all the other cases discussed so far, Rosenstein and Cohen (1998) try to solve the SGP through a bottom-up process "from the perception to the elaboration of the

[7] For an approach close to Vogt's and that incurs the same problems see Baillie (2004).

language through the symbolic thought" (Rosenstein and Cohen 1998, p. 20).[8] Unlike the others, their strategy for solving the SGP is based on three components:

1. a method for the organization of the perceptual data, called the *method of delays* or *delays-space embedding*, which apparently allows the AA to store perceptual data without using extrinsic criteria, thus avoiding any semantical commitment;
2. a *predictive semantics*; and
3. an unsupervised learning process, which allows the elaboration of an autonomous semantics.

Consider an example adapted from Rosenstein and Cohen (1999b). ROS is an AA that can move around in a laboratory. It is provided with sensors through which it can perceive its external environment. ROS is able to assess the distance between itself and the objects situated in the external environment. It registers distances at regular time intervals and plots distance and time values on a Cartesian coordinate system, with time on the x-axis and distances on the y-axis. Suppose ROS encounters an object. ROS does not know whether it is approaching the object but its sensor registers that, at time *t*, ROS is at 2000mm from the object, at *t+1* ROS is at 2015mm from the object, and so forth. From these data, we and presumably ROS can deduce that it is moving away from the object. According to Rosenstein and Cohen, an AA like ROS can "know" the consequences of similar actions through the Cartesian representation of the data concerning those actions. The AA envisioned by Rosenstein and Cohen identifies the meaning of its symbols with the outcome of its actions through a Cartesian representation of its perceived data. Since the data plotted on a Cartesian coordinate system define an action, the AA associates with that particular "Cartesian map" the meaning of the corresponding action.

Suppose now that a population of AAs like ROS interact in a simulated environment adopting several strategies for pursuit or avoidance.

[8] The strategy is developed in several papers, see Oates, et al. (1998a), Oates, et. al (1998b), Rosenstein and Cohen (1999a), Rosenstein and Cohen (1999b), Sebastiani, et al. (1999), Cohen, et al. (2002), Cohen (2002), Firoiu and Cohen (2002).

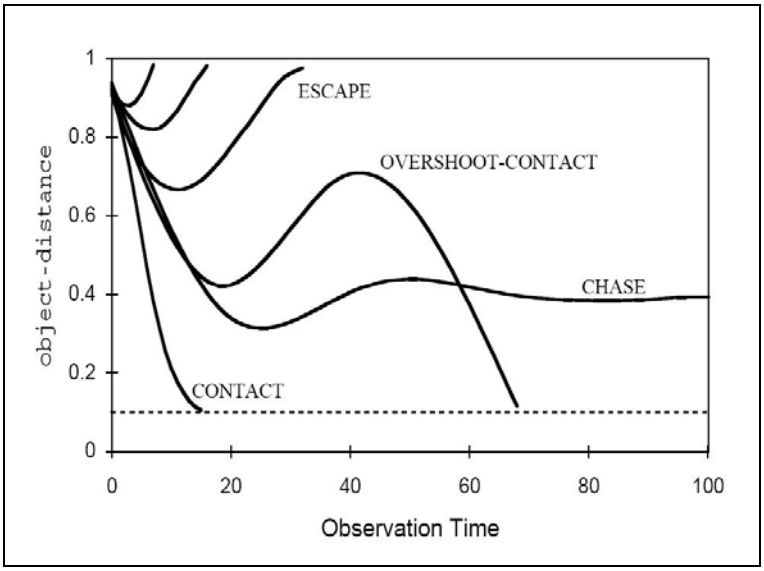

Figure 3 "Cluster prototypes for 100 interactions in the pursuit/avoidance simulator." (Rosenstein and Cohen 1998, p. 21).

Figure 3 shows the six prototypes derived from 100 agent interactions with randomly chosen strategies. According to Rosenstein and Cohen, the categories "chase", "contact", "escape" etc. acquire their meanings in terms of the predictions that each of them enables the AA to make.

As one can see from Figure 3, the actions that have similar outcomes/meaning also have the same Cartesian representation. Rosenstein and Cohen call *natural clustering* this feature of the Cartesian representation. They maintain that, thanks to natural clustering, an AA can elaborate categorical representations of its actions and that, since the Cartesian map already associates action outcomes with meanings, the categories too have a meaning and thus they are semantically founded. Once some initial categories are semantically grounded, the AA can start to elaborate its conceptual representations. The latter are the result of both a comparison of similar categorical representations and of an abstraction of features shared by them. Like the categorical representations on which they are based, the conceptual representations too are semantically grounded. The "artificial" semantics built in this way can grow autonomously, through the interactions of the AA with its environment, until the process allows the AA to predict the outcome of its actions while it is performing them. The prediction is achieved using a learning algorithm. When an AA has a new experience, the algorithm compares the new actions with the ones already represented by previous Cartesian representations, in order to identify and correlate similar patterns. If the AA can find the category of the corresponding actions, it can predict

the outcome/meaning of the new action. The correlation between Cartesian representations and outcome/meaning of the actions allows the AA to elaborate a *predictive semantics*.

It seems that the SGP is solved without using any external or pre-semantical criteria. Apparently, the only parameter used for the initial categorization of an AA's actions is time, and this cannot be defined as an external parameter, since it is connected with the execution of the actions (Rosenstein and Cohen 1998).

The appearance, however, is misleading. For it is the Cartesian coordinate system, its plotting procedures and symbolic conventions used by the AA that constitute the pivotal, semantic framework allowing the elaboration of an initial semantics by an AA like ROS. But clearly this "Cartesian" semantic framework is entirely extraneous to the AA, either being presumed to be there (innatism) or, more realistically, having been superimposed by the programmer. Rosenstein and Cohen seem to consider an AA mapping of its actions on some Cartesian coordinates as some sort of *spontaneous* representation of the perceptual data by the AA itself. However, the very interpretation of the data, provided by the actions, as information of such and such a kind on a Cartesian coordinate system is, by itself, a crucial semantic step, based on extrinsic criteria. Obviously, the system does not satisfy the semantical commitment criterion, and the approach fails to solve the SGP.

With the temporal delays method, we conclude the part of this paper dedicated to the semi-representationalist approach to the SGP. Again, none of the hypotheses discussed appears to provide a valid solution for the SGP. In the next section, we shall see what happens when representationalism is discarded in favour of a non-representationalist approach to the SGP.

6. The Non-representationalist Approach

The roots of a non-representationalist approach to the SGP may be dated to the criticisms made by Brooks (1990) and Brooks (1991) of the classic concept of representation. Brooks argues that intelligent behaviour can be the outcome of interactions between an *embodied* and *situated*[9] AA and its environment and that, for

[9] An AA is embodied if it is implemented in a physical structure through which it can have direct experience of its surrounding world. The same AA is also situated if it is placed in a dynamic environment with which it can interact.

this purpose, symbolic representations are not necessary, only sensorimotor couplings. This is what Brooks (1991) calls the *Physical Grounding Hypothesis*.

In order to explore the construction of physically grounded systems, Brooks has developed a computational architecture known as the *subsumption architecture*, which "enables us to tightly connect perception to action, embedding robots correctly in the world." (Brooks 1990, p. 5). The details of Brooks' subsumption architecture are well known and there is no need to summarise them here. What is worth emphasizing is that, since a subsumption architecture allows an AA to avoid any elaboration of explicit representations, within this paradigm one may argue that the SGP is solved in the sense that it is entirely avoided: if there are no symbolical representations to ground, there is no symbol grounding problem to be solved.

However, the SGP is merely postponed rather than avoided: an AA implementing a subsumption architecture may not need to deal with the SGP initially, in order to deal successfully with its environment; but if it is to develop even an elementary protolanguage and some higher cognitive capacities, it will have to be able to manipulate some symbols, but then the question of their semantic grounding presents itself anew. This is the problem addressed by the following two strategies.

6.1. A Communication-based Model for the Solution of the SGP

Billard and Dautenhahn (1999) propose a communication-based approach to the SGP that can be interpreted as steering a middle course between the strategies advocated by Vogt (2002a) and by Varshavskaya (2002) (see next section).

The topic of their research is AAs' social skills in learning, communicating and imitating. They investigate grounding and use of communication through simulations within a group of AAs. In this context, we find their proposal on how to approach the SGP.

The experimental scenario consists of nine AAs interacting in the same environment and sharing a common set of perceptions. The AAs have short-term memory, and they are able to move around, communicate with each other and describe their internal and external perceptions. Their task is to learn a common language through a simple imitation game. In the experiment, the AAs are expected to learn a vocabulary to differentiate between coloured patches and to describe their locations in terms of distance and orientation, relative to a "home point". "The vocabulary is transmitted from a teacher agent, *which has a complete knowledge of*

the vocabulary from start [emphasis added], to eight learner agents, which have no knowledge of the vocabulary at the start of the experiments" (Billard and Dautenhahn 1999, p. 414-415). Transmission of the vocabulary from teacher to learner occurs as part of an imitative strategy. Learning the vocabulary, or grounding of the teacher's signals in the learner's sensor-actuator states, results from an association process across all the learner's sensor-actuator, thanks to a Dynamic Recurrent Associative Memory Architecture (DRAMA). DRAMA has a "considerable facility for conditional associative learning, including an efficient short-term memory for sequences and combinations, and an ability to easily and rapidly produce new combinations", (Billard and Dautenhahn 1999, p. 413).

According to Billard and Dautenhahn, the experiment indicates a valuable strategy for overcoming the SGP; "Our work showed the importance of behavioural capacities alongside cognitive ones for addressing the symbol grounding problem." (Billard and Dautenhahn 1999, p. 429). However, it is evident that the validity of their proposal is undermined by three problems. First, the learning AAs are endowed with semantic resources (such as their DRAMA) whose presence is merely presupposed without any further justification (innatism). Note also that in this context there is a reliance on neural networks, which incurs the same problems highlighted in section 4.1. Second, the learning AAs acquire a pre-established, complete language from an external source (externalism); they do not develop it by themselves through their mutual communications and their interactions with their environment. Third, the external source-teacher is merely assumed to have full knowledge of the language and the semantics involved. This is another form of "innatism" utterly unjustified in connection with the SGP. The hard question is how the teacher develops its language in the first place. This is the SGP, but to this Billard and Dautenhahn provide no answer. The result is that the strategy begs the question thrice and cannot be considered a valid solution of the grounding problem.

6.2. A Behaviour-based Model for the Solution of the SGP

Following Brooks (1991), Varshavskaya (2002) argues that the development of semantic capacities in an AA could be modelled on the development of linguistic capacities in children. Theories of language acquisition appear to show that children acquire linguistic skills by using a language as a tool with which to interact with their environment and other agents, in order to satisfy their needs and achieve their goals.

Accordingly, Varshavskaya supports a *pragmatic* interpretation of language acquisition in AA whereby "[l]anguage is not viewed as a denotational symbolic system for reference to objects and relationships between them, as much as a tool for communicating intentions. The utterance is a way to manipulate the environment through the beliefs and actions of others", (Varshavskaya 2002, p. 149). Language becomes just another form of pragmatic interaction of the AA with its environment and, as such, its semantics does not need representations.

The hypothesis of a representations-free language has been corroborated by some experiments involving a MIT robot known as KISMET (Breazeal 2000).

"KISMET is an expressive robotic head, designed to have a youthful appearance and perceptual and motor capabilities tuned to human communication channels. The robot receives visual input from four color CCD cameras and auditory input from a microphone. It performs motor acts such as vocalizations, facial expressions, posture changes, as well as gaze direction and head orientation." (Varshavskaya 2002, p. 151). The experiments show that KISMET can learn from its trainer to use symbols and to develop protolinguistic behaviours. Varshavskaya states that, in so doing, KISMET has made the first steps towards the development of much more complex linguistic capacities.

Learning to communicate with the teacher using a shared semantics is for KISMET part of the more general task of learning how to interact with, and manipulate, its environment. KISMET has motivational (see next section) and behavioural systems and a set of vocal behaviours, regulatory drives, and learning algorithms, which together constitute its *protolanguage* module. Protolanguage refers here to the "pre-grammatical" time of the development of a language – the babbling time in children – which allows the development of the articulation of sounds in the first months of life. To KISMET, protolanguage provides the means to ground the development of its linguistics capacities.

KISMET is an autonomous AA, with its own goals and strategies, which cause it to implement specific behaviours in order to satisfy its "necessities". Its "motivations" make it execute its tasks. These motivations are provided by a set of homeostatic variables, called *drives*, such as the level of engagement with the environment or the intensity of social plays. The drives must be kept within certain bounds in order to maintain KISMET's system in equilibrium. Kismet has "emotions" as well, which are a kind of motivation.

Emotion	Behavior	Proto-linguistic Function
anger, frustration	complain	regulatory
disgust	withdraw	instrumental or regulatory
fear, distress	escape	–
calm	engage	interactional
joy	display pleasure	personal or interactional
sorrow	display sorrow	regulatory or personal
surprise	startle response	–
boredom	seek	–

Figure 4 "The correspondence between KISMET's nonverbal behaviours and protolinguistic functions", (Varshavskaya 2002, p. 153).

KISMET's emotions depend on the evaluations of the perceptual stimuli. When the homeostatic values are off-balance, KISMET can perform a series of actions that allow it to regain a pre-established equilibrium. In these cases, KISMET uses some protoverbal behaviours – it expresses its "emotions" – with which it acts on itself and on the environment in order to resume the balance of the original values.

KISMET can implement protolinguistic behaviours, thanks to the presence of two drives (one for the language and one for the exploration of the environment), an architecture to express protoverbal behaviours and an architecture for the visual apparatus. The language drive allows two behaviours called *Reader* and *Hearer* (see Figure 5) "which interface with KISMET's perceptual system and procure global releasers for vocal behavior" (Varshavskaya 2002, p. 153). There is also a *Speaker* behaviour responsible for sending a speech request over to the robot.

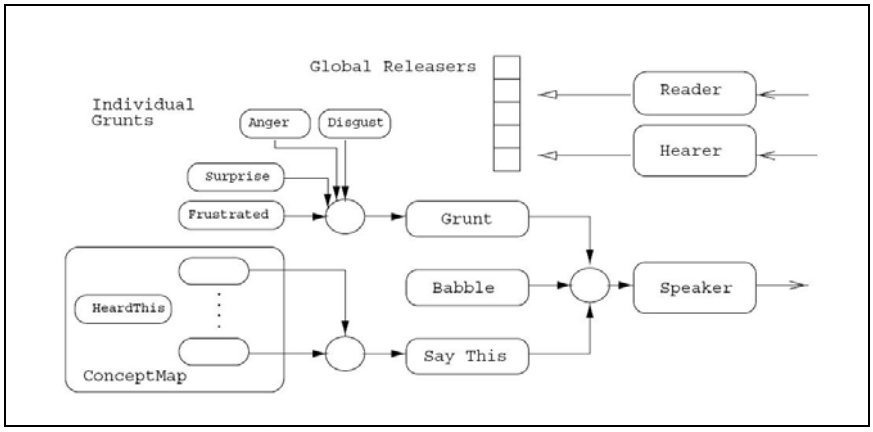

Figure 5 "Overall architecture of KISMET's protoverbal behaviors, where rounded boxes represent instances of behaviors and circles represent connections between behaviors. Connections between HeardThis and individual Concepts are not shown for clarity", (Varshavskaya 2002, p. 154).

The kind of requests depends on the competition between the individual protoverbal behaviours that KISMET can perform. These are in a competitive hierarchy and the one which has the highest position in the hierarchy is executed.

Let us now see, with an example, what the emulation processes are and how they influence KISMET's learning process. Suppose KISMET learns the English word "green". The trainer shows KISMET a green object and at the same time she utters the word "green", while KISMET is observing the green object. Then the trainer hides the green object, which will be shown again only if KISMET will look for it and expresses a vocal request similar to the word "green". If KISMET utters the word "green" in order to request the green object, then KISMET has learned the association between the word and the object, and to use the word according to its meaning. By performing similar tasks KISMET seems to be able to acquire semantical capacities and to develop them without elaborate representations. We shall now see whether this may be sufficient to solve the SGP.

6.2.1. Emulative Learning and the Rejection of Representations

The learning approach adopted by Varshavskaya is intrinsically inadequate to deal with the SGP successfully. For the question concerning the origin of semantic capacities in artificial systems – i.e. how KISMET begins to semanticise in the first place – cannot be addressed by referring to modalities appropriate to human agents, since only in this case it is correct to assume

- a natural and innate predisposition in the agent to acquire a language;

- the existence of an already well-developed language; and
- the presence of a community of speakers, proficient in that language, who can transmit knowledge of that language to new members.

None of these assumptions is justified when an AA is in question, including KISMET. Recall that, in order to solve the SGP, the semantic capacities of the AA must be elaborated by the AA itself autonomously, without begging the question: no innatism or externalism is allowed. Yet, both occur in KISMET's case. KISMET is (innately) endowed with semantic features (recall the presence of a protolanguage) and it (externally) performs an explicitly emulative learning. It associates the symbol 'green' to the green object shown by the trainer, but the initial, semantic relation between 'green' and the green object is pre-established and provided by the trainer herself. As far as the SGP is concerned, teaching KISMET the meaning of 'green' is not very different from uploading a lookup table.

The point may be further clarified by considering the following difficulty: does the symbol 'green' for KISMET refer to the specific green object shown to KISMET by the trainer or does it, instead, name a general feature – the colour of the green object – that KISMET can recognize in that as well as in other similar objects? Suppose we show KISMET several objects, with different shapes but all having the property of being green. Among these objects, there is also the green object that KISMET already knows. If one asks KISMET to recognize a green object it will recognize only the green object it has seen before. This is so because KISMET does not name classes of objects, e.g. all the green objects. Instead, it has symbols that name their referents rigidly, as if they were their proper names. For KISMET, the green object will not *be green*, it will *be called 'green'*, in the same sense in which a black dog may be called "Blackie". This follows from KISMET's non-representationalist elaborations. KISMET's semantics can grow as much as the emulative learning process externally superimposed by the trainer allows, but the absence of representations means that Kismet will not develop any categorical framework in the sense required to solve the SGP. Lacking representations, KISMET is unable to connect a symbol to a category of data.

7. Conclusion

In this paper, we have outlined the SGP, defended the zero semantical commitment condition (Z condition) as the requirement that must be satisfied by a strategy in order

to provide a valid solution of the SGP, and then reviewed eight strategies developed for solving the SGP in the last fifteen years. We have organised them into three approaches: representationalism, semi-representationalism and non-representationalism. In the course of the review, we have shown that all eight strategies are semantically committed, and hence that none provides a valid solution of the SGP.

The positive lesson that can be learnt from the reviewed research is that (the semantic capacity to generate) representations cannot be presupposed without begging the question. Yet abandoning any reference to representations means accepting a dramatic limit to what an AA may be able to do semantically, since the development of even the simplest abstract category becomes impossible. So it seems that a valid solution of the SGP will need to combine at least the following features:

1. a bottom-up, sensorimotor approach to the grounding problem;
2. a top-down feedback approach that allows the harmonization of top level grounded symbols and bottom level, sensorimotor interactions with the environment;
3. the availability of some sort of representational capacities in the AA;
4. the availability of some sort of categorical/abstracting capacities in the AA;
5. the availability of some sort of communication capacities among AAs in order to ground the symbols diachronically and avoid the Wittgensteinian problem of a "private language";
6. an evolutionary approach in the development of (1)-(5);
7. the satisfaction of the Z condition in the development of (1)-(6).

Whether all this may be possible even in principle is an entirely different issue, whose exploration has been left to a second stage of this research.[10]

[10] Research for this paper has been supported by a grant awarded to Mariarosaria Taddeo by the Università degli Studi di Bari. We are very grateful to Gian Maria Greco, Gianluca Paronitti and Matteo Turilli for their many and very helpful comments on previous drafts of this article. For all Italian legal requirements, Mariarosaria Taddeo must be considered the first author of the article and the author of sections 2, 3, 4, 5 and 6.

References

J. C. Baillie, “Grounding Symbols in Perception with two Interacting Autonomous Robots” in *Proceedings of the Fourth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems 117* Genoa, Italy, 2004, pp. 107-110.

A. Billard and K. Dautenhahn, “Experiments in Learning by Imitation – Grounding and Use of Communication in Robotic Agents”, *Adaptive Behaviour*, 7, pp. 411-434, 1999.

C. Breazeal, “Sociable Machines: Expressive Social Exchange Between Humans and Robots”. Sc.D. dissertation, Department of Electrical Engineering and Computer Science, MIT, 2000.

R. A. Brooks, “Elephants Don’t Play Chess”, *Robotics and Autonomous Systems*, 6, pp. 3–15, 1990.

R. A Brooks, “Intelligence Without Representation”, *Artificial Intelligence Journal*, 47, pp. 139–159, 1991.

A. Cangelosi, “Evolution of Communication and Language Using Signals, Symbols and Words”, *IEEE Transaction in Evolution Computation*, 5, pp. 93-101, 2001.

A. Cangelosi, A. Greco and S. Harnad, “From Robotic Toil to Symbolic Theft: Grounding Transfer from Entry-Level to Higher-Level Categories”, *Connection Science*, 12, pp. 143-162, 2000.

A. Cangelosi, A. Greco and S. Harnad, “Symbol Grounding and the Symbolic Theft Hypothesis”, in *Simulating the Evolution of Language*, A. Cangelosi and D. Parisi, Eds., London, Springer, 2002, pp.191-210.

A. Cangelosi and S. Harnad, “The Adaptive Advantage of Symbolic Theft over Sensorimotor Toil: Grounding Language Perceptual Categories”, *Evolution of Communication, special issue on Grounding Language*, 4, pp. 117-142, 2001.

D. J. Chalmers, “Subsymbolic Computation and the Chinese Room”, in *The Symbolic and Connectionist Paradigms: Closing the Gap*, J. Dinsmore, Ed., Hillsdale: Lawrence Erlbaum, 1992, pp. 25-48.

P. R. Cohen, C. Sutton and B. Burns, “Learning Effects of Robot Actions using Temporal Associations”, in *2nd International Conference on Development and Learning*, 2002, pp. 90-101.

D. Cole, “The Chinese Room Argument”, *The Stanford Encyclopedia of Philosophy (Fall 2004 Edition)*, Edward N. Zalta Ed., URL = http://plato.stanford.edu/archives/fall2004/entries/chinese-room/.

S. Coradeschi and A. Saffioti, “An Introduction to the Anchoring Problem”, *Robotics and Autonomous Systems*, 43, pp. 85-96, 2003.

P. Davidsson, “Toward a General Solution to the Symbol Grounding Problem: Combining Machine Learning and Computer Vision“, in *AAAI Fall Symposium Series, Machine Learning in Computer Vision: What, Why and How?*, 1993, pp. 157-161.

G. Dorffner and E. Prem, “Connectionism, Symbol Grounding, and Autonomous Agents”, in *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society*, 1993, pp. 144-148.

L. Floridi, “Open Problems in the Philosophy of Information”, *Metaphilosophy*, 35, pp. 554-582, 2004.

L. Firoiu and P. R. Cohen, “Segmenting Time Series with a Hybrid Neural Networks - Hidden Markov Model”, in *The Eighteenth National Conference on Artificial Intelligence*, 2002, pp. 247-252.

S. Harnad, (Ed.). *Categorical Perception: the Groundwork of Cognition*, New York: Cambridge University Press, 1987, pp. 287-300.

S. Harnad, “The Symbol Grounding Problem”, *Physica* D, pp. 335-346, 1990.

S. Harnad, “Symbol Grounding in an Empirical Problem: Neural Nets are just a Candidate Component”, in *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society*, 1993.

S. Harnad, “Grounding Symbols in the Analog World with Neural Nets – a Hybrid Model”, *Psychology*, 12, pp. 12-78, 2001.

S. Harnad, “Symbol Grounding and the Origin of Language”, in *Computationalism: New Directions*, M. Scheutz, Ed., MIT Press, 2002, pp. 143-158.

S. Harnad, “The Symbol Grounding Problem”, in *Encyclopedia of Cognitive Science*, London: Nature Publishing Group/Macmillan, 2003.

J. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1975.

M. Mayo, “Symbol Grounding and its Implication for Artificial Intelligence”, in *Twenty-Sixth Australian Computer Science Conference*, 2003, pp. 55-60.

T. Oates, P. R. Cohen and C. Durfee, “Efficient Mining of Statistical Dependencies“, in *Seventh International Workshop on Artificial Intelligence and Statistics*, 1998a, pp. 133-141.

T. Oates, D. Jensen and P. R. Cohen, “Discovering Rules for Clustering and Predicting Asynchronous Events”, in *Predicting the Future: AI Approaches to Time Series Workshop*,1998b, pp. 73-79.

E. Prem, “Symbol Grounding and Transcendental Logic”, in L. Niklasson and M. Boden, Eds., *Current Trends in Connectionism*, Hillsdale, NJ: Lawrence Erlbaum, 1995a, pp. 271-282.

E. Prem “Dynamic Symbol Grounding, State Construction, and the Problem of Teleology”, in J. Mira and F. Sandoval, Eds., *From Natural to Artificial Neural Computation, Proceedings of the International Workshop on Artificial Neural Networks*, Malaga-Torremolinos, Spain, June. Springer, LNCS 930, 1995b, pp. 619-626.

E. Prem, “Grounding and the Entailment Structure in Robots and Artificial Life”, in F. Moran et al., Eds., *Advances in Artificial Life, Proceedings of the Third European Conference on Artificial Life*, Granada, Spain, Springer, 1995c.

C. S. Peirce, *Collected Papers*, Vol. I-VIII, Cambridge MA: Harvard University Press, 1932-1958.

M. T. Rosenstein and P. R Cohen, “Symbol Grounding With Delay Coordinates”, in *AAAI Technical Report WS-98-06, The Grounding of Word Meaning: Data and Models*, 1998, pp. 20-21.

M. Rosenstein and P. R. Cohen, “Continuous Categories for a Mobile Robot”, in *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 1999a, pp. 634-640.

M. Rosenstein and P. R. Cohen, “Concepts from Time Series”, in *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1999b, pp. 739-745.

J. Searle, “Minds, Brains, and Programs”, *Behavioral and Brain Sciences*, 3, pp. 417-458,1980.

P. Sebastiani, M. Ramoni and P. R. Cohen, “Unsupervised Classification of Sensory Input in a Mobile Robot”, in *IJCAI-99 Workshop on Sequence Learning*, 1999, pp. 23-28.

N. E. Sharkey and S. A. Jackson, “Three Horns of the Representational Trilemma”, in *Symbol Processing and Connectionist Models for Artificial Intelligence and*

Cognitive Modeling: Steps towards Integration, V. Honavar and L. Uhr, Eds., Cambridge, MA: Academic Press, 1994, pp. 155-189.

L. Steels and P. Vogt, "Grounding Adaptive Language Games in Robotic Agents", in *Proceedings of Fourth European Conference on Artificial Life*, 1997, pp. 474-482.

P. Slusarz and R. Sun, "The Interaction of Explicit and Implicit Learning: an Integrated Model", in *Proceedings of the 23rd Cognitive Science Society Conference*, 2001, pp. 952-957.

R. Sun, "Learning, Action, and Consciousness: a Hybrid Approach towards Modelling Consciousness", *Neural Networks,* special issue on consciousness, 10, pp. 1317-1331, 1997.

R. Sun, "Symbol Grounding: A New Look at an Old Idea", *Philosophical Psychology*, 13, pp. 149-172, 2000.

R. Sun, "Computation, Reduction, and Teleology of Consciousness", *Cognitive Systems Research*, 1, pp. 241-249, 2001.

R. Sun, E. Merrill and T. Peterson, "From Implicit Skills to Explicit Knowledge: a Bottom-up Model of Skill Learning", *Cognitive Science*, 25, pp. 203-244, 2001.

R. Sun and T. Peterson, "Some Experiments with a Hybrid Model for Learning Sequential Decision Making", *Information Science*, 111, pp. 83-107, 1998.

R. Sun and X. Zhang, "Top-down versus Bottom-up Learning in Skill Acquisition", in *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002, pp. 63-89.

J. L. Taylor and S.A. Burgess, "Steve Austin versus the Symbol Grounding Problem", in *Proceedings of Selected Papers from the Computers and Philosophy Conference* (CAP2003), Canberra, Australia, *Conferences in Research and Practice in Information Technology*, 37, 2004, pp. 21-25.

P. Varshavskaya, 2002, "Behavior-Based Early Language Development on a Humanoid Robot", in *Proceedings of the Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, 94, pp. 149-158.

P. Vogt, "Perceptual Grounding in Robots", in *Proceedings of the 6th European Workshop on Learning Robots 1997. Lecture Notes on Artificial Intelligence*, 1545, 1997, pp. 36-45.

P. Vogt, "Bootstrapping Grounded Symbols by Minimal Autonomous Robots", *Evolution of Communication*, 4, 2000, pp. 89-118.

P. Vogt, “The Physical Symbol Grounding Problem”, *Cognitive Systems Research* 3, 2002a, pp. 429-457.

P. Vogt, “Anchoring Symbols to Sensorimotor Control“, in *Proceedings of Belgian/Netherlands Artificial Intelligence Conference BNAIC’02*, 2002b.

P. Vogt and H. Coumans, “Exploring the Impact of Contextual Input on the Evolution of Word-Meaning”, in *Proceedings of the Seventh International Conference of The Society for Adaptive Behavior*, 2002.

P. Vogt and H. Coumans, “Investigating Social Interaction Strategies for Bootstrapping Lexicon Development“. *Journal of Artificial Societies and Social Simulation*, 6, 2003.

T. Ziemke, “Rethinking Grounding”, in *Understanding Representation in the Cognitive Sciences*, A. Riegler, M. Peschl and A. von Stein, Eds., New York: Plenum Press,1999, pp.177-190.