A Cognitive Architecture for Robot Self-Consciousness

Antonio Chella\textsuperscript{a,1}, Marcello Frixione\textsuperscript{b,2} and Salvatore Gaglio\textsuperscript{a,3}

\textsuperscript{a} Dipartimento di Ingegneria Informatica, Università di Palermo, Viale delle Scienze, 90128 Palermo, Italy

\textsuperscript{b} Dipartimento di Scienze della Comunicazione, Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (Salerno), Italy

Summary

Objective:

One of the major topics towards robot consciousness is to give a robot the capabilities of self-consciousness. We propose that robot self-consciousness is based on higher order perception of the robot, in the sense that first order robot perception is the immediate perception of the outer world, while higher order perception is the perception of the inner world of the robot.

Methods and Material:

We refer to a robot cognitive architecture that has been developed during almost ten years at the RoboticsLab of the University of Palermo. The architecture is organized in three computational areas. The subconceptual area is concerned with the low level processing of perceptual data coming from the

\textsuperscript{1} Corresponding author. E-mail: chella@unipa.it. Phone: +39 091 23842515. Fax: +39 091 6598043

\textsuperscript{2} E-mail: frix@dist.unige.it

\textsuperscript{3} E-mail: gaglio@unipa.it

Preprint submitted to Elsevier 5 July 2008
sensors. In the linguistic area, representation and processing are based on a logic formalism. In the conceptual area, the data coming from the subconceptual area are organized in conceptual categories.

**Results:**

To model higher order perceptions in self reflective agents, we introduce the notion of second-order points in conceptual space. Each point in this space corresponds to a self reflective agent, i.e., the robot itself, persons, and other robots with introspective capabilities.

**Conclusions:**

The described model of robot self-consciousness, although effective, highlights open problems from the point of view of the computational requirements of the current state-of-art computer systems. Some future works that lets the robot to summarize its own past experiences should be investigated.

**Keywords:**

Self-consciousness; Machine consciousness; Conceptual spaces; Cognitive systems.

1 **Introduction**

In recent years, there has been an increasing interest towards consciousness and self-consciousness. Following this interest, computational models of machine consciousness for autonomous robots have been proposed and discussed, see Chella and Manzotti [1] for a review.

One of the major topics towards robot consciousness is to give a robot the capabilities of self-consciousness, i.e., to reflect about itself, its own perceptions and actions during its operating life. The robot sense of self grows up from the content of the agent perceptions, recalls, actions, reflections and so on in a coherent life long narrative.

A first theoretical founded attempt to give self reflection capabilities to an artificial reasoning system is described in the seminal paper of Weyhrauch [2].
Fig. 1. The computational areas of the proposed architecture.

Weyhrauch proposed the FOL system based on formal logic able to perform logic inference and to reflect about its own inferences.

McCarthy [3] stresses the fact that a robot needs the ability to observe its own mental states. He proposes the mental situation calculus as a formalism that extends the situation calculus in order to represent mental situations and actions.

Minsky [4] describes an hypothetical system based on several interacting agents at different levels, in which the tasks of higher levels agents is the self-reflective processing. A first attempt to implement the system proposed by Minsky in a simulated world is described in Singh and Minsky [5].


McDermott [7] makes a distinction between normal access to the output of a computational module and and introspective access to the same module. The first one is related with the output related with the processing algorithms of the module, while the second one is related with the higher-order access inside of the processing of the module according to the self model. He also discusses the relationships between higher-order access and phenomenology, in the line of higher-order theories of consciousness (see, e.g., Carruthers [8]).

In this paper we propose a model of robot self-consciousness based on higher order perceptions of the robot during time, in the sense that first order robot perceptions are the immediate perceptions of the outer world of a self reflective agent, while higher order perceptions are the perceptions during time of the inner world of the agent. A preliminary version of the model is described in Chella and Gaglio [9].
2 The Cognitive Architecture

We refer to a robot cognitive architecture that has been developed during almost ten years at the RoboticsLab of the University of Palermo. The architecture has been successfully adopted in different robotics contexts: robot vision [10,11], action planning [12,13], symbol anchoring [14], imitation learning [15], control for robot museum guide [16].

The architecture is organized in three computational areas. Fig. 1 schematically shows the relations among them. The subconceptual area is concerned with the low level processing of perceptual data coming from the sensors. We call it “subconceptual” because here information is not yet organized in terms of conceptual structures and categories.

In the linguistic area, representation and processing are based on a logic formalism, i.e., first order logic. In the conceptual area, the data coming from the subconceptual area are organized in conceptual categories, which are still independent from any linguistic characterization. The symbols in the linguistic area are anchored to sensory data by mapping them on the representations in the conceptual area.

2.1 Conceptual Area

Representations in the conceptual area are couched in terms of a conceptual space (Gärdenfors [17]) that provides a principled way for relating high level, linguistic formalisms on the one hand, with low level, unstructured representation of data on the other. A conceptual space CS is a metric space whose dimensions are related to the quantities processed in the subconceptual area. Dimensions do not depend on any specific linguistic description. In this sense, a conceptual space comes before any symbolic-propositional characterization of cognitive phenomena.

The term knoxel denotes a point in a conceptual space. From the mathematical point of view, a knoxel $k$ is a vector in CS; from the conceptual point of view, it is the epistemologically basic element at the considered level of analysis. In the case of static scenes [10], a knoxel coincides with a 3D primitive shape, described in terms of some constructive solid geometry (CSG) schema.

In particular, following Pentland [18] we adopt superquadrics as knoxel primitive shapes. In brief, superquadrics are geometric shapes derived from the quadric parametric equation with the trigonometric functions raised to two real exponents.
The inside/outside function of the superquadric in implicit form is:

\[
F(x, y, z) = \left( \frac{x}{a_x} \right)^{\frac{2}{\varepsilon_1}} + \left( \frac{y}{a_y} \right)^{\frac{2}{\varepsilon_2}} + \left( \frac{z}{a_z} \right)^{\frac{2}{\varepsilon_1}} \tag{1}
\]

where the parameters \(a_x, a_y, a_z\) are the lengths of the superquadric axes and the exponents \(\varepsilon_1, \varepsilon_2\), called form factors, are responsible for the shape’s form: \(\varepsilon_1\) acts in terms of the longitude, and \(\varepsilon_2\) in terms of the latitude of the object’s surface. (1) returns a value equal to 1 when the point \((x, y, z)\) is a superquadric boundary point, a value less than 1 when it is an inside point, and a value greater than 1 when it is an outside point. The superquadric takes on a squared shape when the form factors values are less than 1, and it takes on a rounded shape when the form factors values are near 1. Fig. 2 shows the shapes of a superquadric when varying the form factors \((\varepsilon_1, \varepsilon_2)\).

The mathematical representation of a knoxel \(k\) is obtained from Eq. 1 by adding the three center coordinates \(p_x, p_y, p_z\) and the three orientation parameters \(\varphi, \vartheta, \psi\) of the superquadric.
A metric function is defined in CS, according to which similar entities correspond to neighboring knoxels in the space. Such metric function may not be explicitly defined. Rather, it may be implicitly computed, e.g., by means of suitable neural networks (as it is the case of the approach we developed). The metric in CS introduces a measure of the degree of “typicality” of an individual as a member of a category. Convex clusters of knoxels are good candidates for the interpretation of linguistic symbols expressing “natural” categories, according to Gärdenfors [17].

In order to represent dynamic scenes, we adopted an intrinsically dynamic conceptual space [11]. The main assumption behind such a dynamic CS is that simple motions are categorized in their wholeness, and not as sequences of static frames. According to this hypothesis, every knoxel now corresponds to a simple motion of a superquadric.

Formally, the knoxel $\mathbf{k}$ of the dynamic CS can be decomposed in a set of components $m_i(t)$, each of them associated with a degree of freedom of the
moving superquadric. In other words:

\[ k = [x_1(t), x_2(t), \ldots, x_{11}(t)] \]  

(2)

where \( x_1(t) = a_x(t), \ x_2(t) = a_y(t) \), and so on. In turn, each motion \( x_i(t) \) may be considered as the result of the superimposition of a set of elementary motions \( f_j^i(t) \):

\[ x_i(t) = \sum_j X^i_j f_j^i(t) \]  

(3)

In this way, it is possible to choose a set of basis functions \( f_j^i(t) \), in terms of which any simple motion can be expressed. Such functions can be associated to the axes of the dynamic conceptual space as its dimensions. Therefore, from the mathematical point of view, the resulting CS is a functional space.

In the domain under investigation, the chosen set of basis functions are the first low frequency harmonics, according to the well-known Discrete Fourier Transform, see, e.g., Oppenheim and Shafer [19]. By a suitable composition of the trigonometric functions of all of the geometric parameters, the overall motion of a 3D primitive is represented as a point in the functional space.

A single knoxel in the dynamic CS now describes a simple motion, i.e., the motion of a primitive shape. A situation is a motion of all the objects in a scene approximated by more than one primitive shape. A situation is represented in the CS by the set of knoxels corresponding to the motions of its components. For example, the motion of the robot with respect to all obstacle may be represented in CS by the set of the knoxels corresponding to its components, as in Fig. 3(a) where \( k_a \) corresponds to the robot, and \( k_b \) corresponds to an obstacle object:

\[ CS = \{k_a, k_b\} \]  

(4)

Note that in a situation the simple motions of their components occur simultaneously. To consider the composition of several simple or composite motions arranged according to some temporal relation (e.g., a sequence), the notion of action is introduced. An action corresponds to a series of different configurations of knoxels in the conceptual space. In the transition between two subsequent different configurations, there is a change of at least one of the knoxels in the CS which is the consequence of a change in the motion of the corresponding superquadrics. We call scattering such a transition from one knoxel to another. It corresponds to a discontinuity in time, and is associated with an instantaneous event.
Fig. 4. A pictorial representation of a scattering between two situations in the conceptual space.

An example of action performed by the robot is the Avoid action. Let us consider Fig. 3: the robot at first moves to the left in order to avoid the obstacle and then it turns on the right in order to come back to its trajectory. In the CS representation, this amounts to say that knoxel $k_a$ (i.e., the robot oriented towards the obstacle) is replaced by knoxel $k'_a$ (i.e., the robot oriented to the left). The new CS’ configuration is:

$$CS' = \{k'_a, k_b\}.$$  \hspace{1cm} (5)

The occurred scattering may be described by the ordered set of the two CS configurations, before and after the scattering:

$$(CS, CS') \equiv (\{k_a, k_b\}, \{k'_a, k_b\}).$$  \hspace{1cm} (6)

Fig. 4 shows a pictorial representation of the dynamic conceptual space when the scattering occurs.

2.2 Linguistic Area

In the linguistic area, the representation of situations and actions is based on a high level, logic oriented formalism. The linguistic area acts as a sort of long term memory, in the sense that it is a semantic network of symbols and their relationships related with the robot perceptions and actions. The linguistic area also performs inferences of symbolic nature. In the current implementation, the linguistic area is based on a hybrid KB in the KL-ONE tradition [20]. A hybrid formalism in this sense is constituted by two different components: a
In the domain of robot actions, the terminological component contains the description of relevant concepts such as situations, actions, time instants, and so on. The assertional component stores the assertions describing specific situations and actions. Fig. 5 shows a fragment of the terminological knowledge base. In the upper part of the figure some highly general concept is represented. In the lower part, the Avoid concept is shown, as an example of the description of an Action in the terminological KB.

In general, we assume that the description of the concepts in the symbolic KB is not completely exhaustive. We symbolically represent only that information that is necessary for inferences.

The assertional component contains facts expressed as assertions in a predicative language, in which the concepts of the terminological components correspond to one argument predicates, and the roles (e.g. precond, part_of) correspond to two argument relations. For example, the following predicates describe that the instance av1 of the action Avoid has as a precondition the instance bp1 of the situation Blocked_path and it has as an effect the situation Free_path:

Avoid(av1)
Blocked_path(bp1)
Free_path(fp1)
precond(av1,bp1)
effect(av1,fp1).
The linguistic area assigns some names (symbols) to the robot perceived entities, describing their structure with a logical-structural language. As a result, all the symbols in the robot linguistic area find their meaning in the conceptual space that is inside the system itself, this way solving the problem of symbol grounding.

3 The Robot Perception Model

A finite agent with bounded resources cannot carry out a one-shot, exhaustive, and uniform analysis of the acquired data within reasonable resource constraints. Some of the acquired data (and of the relations among them) are more relevant than others, and it should be a waste of time and of computational resources to detect true but useless details. In order to avoid the waste of computational resources, the association between symbolic representations and configurations of knoxels in CS is driven by a sequential scanning mechanism that acts as some sort of internal focus of attention, and is inspired by the attentive processes in human vision.

In our architecture, the perception model is based on a focus of attention that selects the relevant aspects of a perceived scene by sequentially scanning the knoxels in the conceptual space. It is crucial in determining which assertions must be added to the linguistic knowledge base: not all true (and possibly useless) assertions are generated, but only those that are judged to be relevant on the basis of the attentive process.

The recognition of a certain component of a situation (a knoxel in CS) will elicit the expectation of other components of the same situation in the scene. In this case, the mechanism seeks for the corresponding knoxels in the current CS configuration. We call this type of expectation synchronic because it refers to a single situation in CS.

The recognition of a certain situation in CS could also elicit the expectation of a scattering in the arrangement of the knoxels in the scene; i.e., the mechanism generates the expectations for another situation in a subsequent CS configuration. We call this expectation diachronic, in the sense that it involves subsequent configurations of CS. Diachronic expectations can be related with the link existing between a situation perceived as the precondition of an action, and the corresponding situation expected as the effect of the action itself. In this way diachronic expectations can prefigure the situation resulting as the outcome of an action.

We take into account two main sources of expectations. On the one side, expectations could be generated on the basis of the structural information stored in
the symbolic knowledge base, as in the previous example of the action *Avoid*. We call these expectations *linguistic*. As soon as a situation is recognized and the situation is the precond of an action, the symbolic description elicit the expectation of the effect situation.

On the other side, expectations could also be generated by purely *Hebbian*, associative mechanisms between situations. Suppose that the robot has learnt that when it sees a person with the arm pointing on the right, it must turn to the right. The system could learn to associate these situations and to perform the related action. We call these expectations *associative*.

Synchronic expectations refer to the same situations of knoxels; diachronic expectations instead involve subsequent configurations of CS, i.e., they involve the effects of the actions, as in the case of the avoid action.

The linguistic and associative expectations let the robot to *anticipate* future interactions with the objects in the environment. The actions performed by the robot in order to interact with a generic object is represented as a sequence of sets of situations in CS. This sequence can be imagined and simulated in the robot’s CS before the interaction really happens in the real world.

### 4 The Self of the Robot

We claim that one of the sources of self-consciousness are *higher order* perceptions of a self-reflective agent. In this sense, *first order* perceptions are the perceptions of the outer world; they generate representations in the agent conceptual space as described in the previous Sects. To model higher order perceptions in self reflective agents, we introduce the notion of *second-order* knoxel. Each second-order knoxel corresponds to a self reflective agent, i.e., the robot itself, persons, and other robots with introspective capabilities.

A second order knoxel at time $t$ corresponding to a self reflective agent now describes the perception of the conceptual space of the agent at time $t - \delta$, i.e., the perception at a previous $\delta$ time of the configuration of knoxels representing the agent itself and the other perceived entities.

For example, let us consider the situation in Fig. 3(a). In this case, the CS of the robot describes the situation made up by the knoxel $k_a$ corresponding to the robot itself and the knoxel $k_b$ corresponding to the obstacle. The second order knoxel $K_a$ at time $t$ represents the robot perception of being in that

---

4 This is a simplification. In general, an agent is described by the set of knoxels corresponding to its parts.
situación at a time \( t - \delta \):

\[
K_a = \left[ \begin{array}{c} k_a \\ k_b \end{array} \right]_{t - \delta}^T.
\] (7)

The second-order knoxel represents the situation at a previous \( \delta \) time describing the robot itself along with the obstacle object.

The outlined procedure may be generalized to consider higher order knoxels: they correspond to the robot’s higher order perceptions of the knoxels of lower order at previous \( \delta \) times. The union of first-order, second-order and higher-order knoxels is at the basis of the robot *self-consciousness*. The robot recursively embeds higher-order models of its own CS’s during its operating life.

Let us consider the *Avoid* action (Fig. 3). In this case, when the robot turns right to start avoiding the obstacle, the third order knoxel \( K'_a \) is:

\[
K'_a = \left[ K_a \{ k'_a \} k_b \right]_{t - \delta}^T.
\] (8)

i.e., \( K'_a \) represents the robot itself along its own evolution in time described by the scattering event at a previous \( \delta \) time.

After the next scattering, the new knoxel \( K''_a \) is:

\[
K''_a = \left[ K'_a \{ k''_a \} k_b \right]_{t - \delta}^T.
\] (9)

i.e., it is a fourth order knoxel that incorporates the third-order knoxel \( K'_a \) and the scattering occurred at a previous \( \delta \) time.

Fig. 6 shows a picture of the the generation of first-order, second-order and higher-order knoxels during the *Avoid* action of the robot.

The robot *self* is therefore generated and supported by the knoxel dynamics, in the sense that the system generates dynamically first-order, second-order and higher-order knoxels during its operations, and this mechanism of generation of higher-order knoxels is responsible for the robot of self-consciousness.

It is to be noticed that higher-order knoxels correspond to meta-predicates in the linguistic area, i.e., symbolic predicates describing the robot perceiving its own situations and actions. These meta-predicates form the basis of the introspective reasoning of the robot [2,3], in the sense that the robot may reason about its own actions in order to generate evaluations about its own
Fig. 6. A pictorial representation of the higher-order CS during the *avoid* action.
performances. Moreover, the robot equipped with the representation of self may generate more complex plans, in the sense that the robot motivations, i.e., its long term goals, may now include also the higher-order knoxels.

There are various analogies between the system described here and that proposed by Weyhrauch [2]. Namely, in both cases there is the possibility of exploiting meta-level representations, and both systems associate some form of analogue representation to the symbolic formalism (the simulation structure in Weyhrauch’s system), the conceptual space in the proposed system.

The described continuous generation of knoxels at different orders poses problems from the computational point of view, in the sense that the physical memory of the robot may be easily filled up with data structures describing the generated knoxels. In the current implementation of the robot, we faced this problem by periodically compressing the data structures corresponding to the older knoxels. In this way the robot does not lose any generated knoxel, but it takes some time to retrieve the oldest ones.

5 Conclusions

The described model of robot self-consciousness highlights several open problems from the point of view of the computational requirements. First of all, the described architecture requires that the 3D reconstruction of the dynamic scenes perceived by the robot during its tasks should be computed in real time and also the corresponding 2D rendering. At the current state of the art in computer vision and computer graphics literature, this requirement may be satisfied only in case of simple scenes with a few objects where all the motions are slow.

Moreover, the generation of the robot self-consciousness requires that the robot should store in the conceptual space at time $t$ all the information at previous $\delta$ times, starting from the beginning of the robot life. This is a hard requirement to be satisfied because of the physical limitations of the robot memory. Some mechanism that lets the robot to summarize its own past experiences should be investigated. One possibility is to make the representations much more blurred as the level grows: higher order CSs could be less detailed than lower level ones.
References


