*Cybernetics And Human Knowing. Vol. 15, nos. 3-4, pp. 15-41*

# Emergence and Downward Causation in Contemporary Artificial Agents:
## Implications for their Autonomy and Some Design Guidelines

*Argyris Arnellos, Thomas Spyrou, John Darzentas*[1]

Contemporary research in artificial environments has marked the need for autonomy in artificial agents. Autonomy has many interpretations in terms of the field within which it is being used and analyzed, but the majority of the researchers in artificial environments are arguing in favor of a strong and life-like notion of autonomy. Departing from this point the main aim of this paper is to examine the possibility of the emergence of autonomy in contemporary artificial agents. The theoretical findings of research in the areas of living and cognitive systems, suggests that the study of autonomous agents should adopt a systemic and emergent perspective for the analysis of the evolutionary development of the notions/properties of autonomy, functionality, intentionality and meaning, as the fundamental and characteristic properties of a natural agent. An analytic indication of the functional emergence of these concepts and properties is provided, based on the characteristics of the more general systemic framework of second-order cybernetic and of the interactivist framework. The notion of emergence is a key concept in such an analysis which in turn provides the ground for the theoretical evaluation of the autonomy of contemporary artificial agents with respect to the functional emergence of their capacities. The fundamental problems for the emergence of genuine autonomy in artificial agents are critically discussed and some design guidelines are provided.

Keywords: Autonomy, Emergence, Functionality, Meaning, Self-organization, Normativity, Agency, Intentionality, Downward Causation

### 1 Autonomy and Agency

Autonomy is a property that is quite easily ascribed to almost every contemporary artificial agent independently of its constructive, developmental, and functional characteristics (see e.g., Hexmoor, Castelfranchi, & Falcone, 2003). However, and not surprisingly though, the same happens with the notion of agency for almost all artificial systems that are able for at least the most basic interaction with their environment, independently of the ways this interaction is realized. As a result, any artificial system that can be observed to exhibit some kind of pro-activeness in terms of taking the initiative, to have a self-ruling and independent ability to perceive its environment, to reason in order to interpret its perceptions, to draw inferences in order to act in its environment, to solve problems, to communicate with other artificial or natural systems and in general, to socialize, is called as *autonomous agent*. Is this

1. Department of Product and Systems Design Engineering - University of the Aegean, 84100, Syros, Greece. Email: arar@aegean.gr, tsp@aegean.gr, idarz@aegean.gr

justified for every artificial system which is the design result of the research being held in the areas of AI, robotics, ALife, multi-agent systems, and so forth.?

The term *autonomous* derives from a Greek composite word (auto = self) and (nomos = law) and although it has many interpretations, it literally means that a system is free of external control and constraint in its action and judgment, that is self-governing, self-steering. In its most basic version it means that a system exists and acts as an independent entity, that is, it self-generates the rules that govern its functioning, or in other terms, it self-generates its self-regulation. In more specific terms, as (Collier, 2002) argues, an autonomous system exhibits a special form of functional organization that contributes to its own governance and uses this governance for its own maintenance in a variable environment. Consequently, the organization of an autonomous system is both the subject and the object of its functionality.

Autonomous systems primarily act in the world for their self-maintenance. The ability to act upon an environment in order to effect a goal-oriented attribution of a certain purpose belongs to an agent and hence, autonomous systems are ultimately agents. Of course, there are several natural systems that can be considered as autonomous agents, but these systems demonstrate different degrees of agency. Concepts such as autonomy and pro-activity, even though "simple" properties such as perception and inference are not a black and white issue, at all. This should be quite expectable for agency, as well, as it does not also come in an all or nothing package, but it has a gradual nature and there are various many different levels of agency in the biological realm. Actually, autonomy drives interaction and profits from it, and as a result enhances the capacity for agency (Arnellos, Spyrou and Darzentas, in press). Agents are not static things, but complex systems interacting with dynamic and complex environments and therefore exhibiting a dynamic nature. Adopting a dynamic and evolutionary view and attempting to project an agent in a future time horizon, one may suggest that there are some dynamic and gradual conceptual and material ingredients that are complexly integrated together to form an agent in various degrees and in various points of evolution.

Considering the above mentioned, and keeping artificial agents in mind, one may conclude that a complete definition of the term *agent* is out of any question and any prospective definition towards this direction should express agency as a capacity with a gradual and evolutionary nature. In order to pursue such definition we try to modify Kampis's (1999) evolutionary definition of agency, which comes as a list of somewhat ad hoc properties of an agent, in a way that the suggested definition is more susceptible to an analysis of its functional characteristics. We suggest that such a strong notion of agency calls for: interactivity – the ability of an agent/cognitive system to perceive and act upon its environment by taking the initiative; intentionality – the ability of an agent to effect a goal-oriented interaction by attributing purposes, beliefs and desires to its actions; and autonomy – which can be characterized as the ability of an agent to function/operate intentionally and interactively based on its own resources.

This definition mentions three fundamental capacities that an agent should exhibit in a somewhat nested way regarding their existence and evolutionary development. Therefore, according to this definition, agency requires interactivity, which in turn implies action upon the environment. This action is not an accidental but an intentional one, as it is a purposeful action directed towards a goal and it is driven by content such as beliefs and desires. Additionally, such an agent exhibits the property of autonomy, as it interacts with the environment in an intentional manner based on its own resources, hence, also based on its internal content. These three properties seem to be quite interdependent, especially when one attempts to understand if it is possible for each one of them to increase qualitatively while the others remain at the same level.

On the other hand, notions such as autonomy, intentionality, beliefs, goal-orientation, cognition, and so forth are philosophically-loaded and quite heavy terms, which bring about controversies in relevant discussions even in the highly theoretical and interdisciplinary scientific domains. Considering that a theoretical and naturalized analysis of an autonomous agent should also be used as an inspiration and a guide for the design of artificial agency, one should be very careful regarding the conceptual burden that may be raised by the theoretical load of the respective terminology. It is not unlikely, at all, that this is one of the reasons that contemporary artificial systems are so easily called autonomous agents. In the desired direction, Collier (1999), from a critical perspective on the domain of complex systems research, suggests that there is a very interesting interdependence between the three above-mentioned properties. Specifically, Collier suggests that there is no function without autonomy, no intentionality without function, and no meaning without intentionality. The interdependence is completed by considering meaning as a prerequisite for the maintenance of a system's autonomy during its purposeful interaction with the environment.

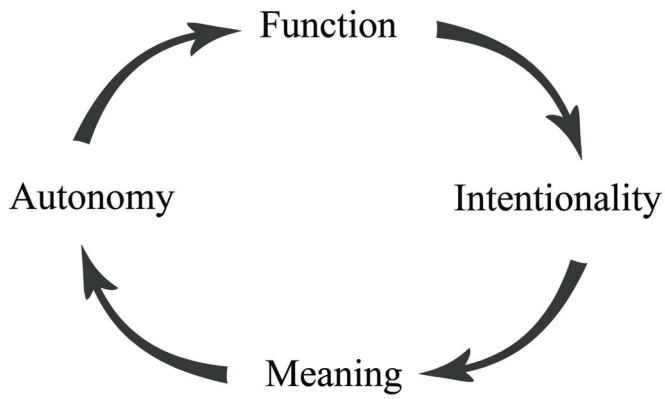



Fig. 1 - Interdependence between autonomy, functionality, intentionality and meaning in an autonomous agent.

At this point some may say that this is just a conceptual interdependence. However, as analyzed below, it is also a theoretical interdependence with a functional grounding, and as such, it sets some interesting constraints in the capacities that contribute to agency and it brings about some requirements in terms of the properties that an agent should exhibit independently of its agential level or in other words, of its level of autonomy and as such, of its cognitive capacity (see Arnellos, Spyrou, Darzentas, 2007b, in press). These properties and their interdependence are characteristics of the strong notion of agency (i.e. the one exhibited by living systems), which is considered as emergent in the functional organization of the living/cognitive system, that is, the autonomous agent. The term *functional* is used here to denote the processes of the network of components that contribute to the autonomy of the agent and particularly, to the maintenance of the autonomous system as a whole (see e.g., Ruiz-Mirazo & Moreno, 2004).

On the other hand, meaning, if it is not to be considered as an ascription of an observer, should be linked with the functional structures of the agent. Hence, meaning should guide the constructive and interactive processes of the functional components of the autonomous system in such a way that these processes maintain and enhance its autonomy. In this perspective, the enhancement of autonomy places certain goals by the autonomous system itself and hence, the intentionality of the system is functionally guiding its behavior through meaning.

At this point, one may ask again the same question, namely if the design results of the broad areas of AI and ALife can be called as autonomous agents. It seems that if one decides to rely upon the analysis so far, he is almost obliged to answer in the negative, although he is not in a very comfortable position with his answer as he still has no means to analytically ground his decision. In the rest of this paper an attempt is made to provide an analysis for such an answer, based on a more analytic indication of the functional formation of the above mentioned concepts and properties that constitute an autonomous agent during its development and evolution. The notion of emergence will prove to be a key concept and the ground for such an answer, as it is directly related to the functionality and the autonomy of an agent.

## 2 Emergent Functionality in Autonomous Agents

One should always keep in mind that in such an autonomous system, intentionality is not reducible to the processing of meanings, nor are the combinations of meanings bringing forth any "aboutness." On the contrary, meaning and its functional substratum are properties that may emerge when an autonomous agent acts intentionally. In other words, an autonomous system may act intentionally if its actions are mediated by meaning. Hence, it appears that for a system to exhibit the capacity of agency, it needs to exhibit the degree of autonomy that will provide for the functionality that is needed, in order to support its intentional and purposeful interaction with the environment, the result of which will emerge new meanings that will further enhance its autonomy. The foundations of such a functional emergence

have been established in the systems-theoretic framework of second-order cybernetics.

*2.1 Emergence through Organizational Closure and Self-Reference*

In the second-order cybernetic epistemology a cognitive system is able to carry out the fundamental actions of distinction and observation. It observes its boundaries and it is thus differentiated from its environment. As the cognitive system is able to observe the distinctions it makes, it is able to refer the result of its actions back to itself. This makes it a self-referential system, providing it with the ability to create new distinctions (actions) based on previous ones, to judge its distinctions, and to increase its complexity by emerging new meanings in order to interact. The self-referential loop can only exist in relation to an environment, but it also disregards the classical system-environment models, which hold that the external control of a cognitive system's adaptation to its environment is replaced by a model of systemic and operational/organizational closure (von Foerster, 1960/2003, 1981).

Due to that closure, the self-reference of an observation emerges meaning inside the cognitive system, which is used as a model for further observations in order to compensate for external complexity. The system which operates on meaning activates only internal functions and structures, (eigenvalues), a set of some stable structures, which are maintained in the functions of the cognitive system's organizational dynamics (Rocha, 1996) and which serve as points of departure for further operations during its interaction with the environment. Indeed, this closure is functional in so far as the effects produced by the cognitive system are the causes for the maintenance of its systemic equilibrium through the emergence of more complex organizations.

With system closure, environmental complexity is based solely on system observations, thus, system reality is observation-based. As von Foerster (1976/2003) argued, the results of an observation do not refer directly to the objects of the real world, but instead, they are the results of recurrent cognitive functions in the structural coupling between the cognitive system and the environment. In particular, von Foerster states that "ontologically, Eigenvalues and objects, and likewise, ontogenetically, stable behavior and the manifestation of a subject's 'grasp' of an object cannot be distinguished" (von Foerster, 1976/2003, p. 266). Thus, each emergent function based on observations is a construction, it is an increase of the organization and cognitive complexity of the agent. This process of emergent increment of order through the internal construction of functional organizations and simultaneous classification of the environment is a process of self-organization (von Foerster, 1960/2003, 1981).

There are two interesting issues at this point. First, self-organizing systems appear to have an emergent functionality which provides the means for self-maintained, self-enhanced and self-regulated organizational dynamics. This functionality originates from a network of processes with a high degree of recursivity that produces and maintains internal invariances in the case of internal and external perturbations. This may be seen as an abstract conception of an autonomous agent, although, as it will be

shown below, it is not (see Fig. 2), but nevertheless, it appears to be a model close enough to many realizations in different biological scales and domains. As such, in the second-order cybernetic framework of autopoiesis (Maturana & Varela, 1980), life is defined as this special kind of basic autonomy (Varela, 1979; Varela & Bourgine, 1992). Actually, Varela says that:

> Autonomous systems are mechanistic (dynamic) systems defined as a unity by their organization. We shall say that autonomous systems are organizationally closed. That is, their organization is characterized by processes such that (1) the processes are related as a network, so that they recursively depend on each other in the generation and realization of the processes themselves, and (2) they constitute the system as a unity recognizable in the space (domain) in which the processes exist. (Varela, 1979, p. 55)

The basis of Varela's conception of autonomy is its active role in the contribution of the self-maintenance of the autopoietic system and especially in the production of its active components, but also in the effective alteration of its boundary conditions in order for the system to be able to maintain its necessary variables in a homeostatic way. What is emphasized in autopoiesis and in self-organization, in general, is the systemic and emergent nature of the whole organism as an autonomous agent. The functionality of the processes of such an autonomous system, as it is described by the two characteristic features in Varela's quotation, can be mapped to an organizational code which executes three functions/operations: a selection of the structural components of the system; their interrelation/correlation in order to emerge a functional whole; and a continuous self-referential control/steering in order to make sure that the respective selection and interrelation are fulfilling the goals of the system.

This code belongs to the designer of the autonomous system, and for the moment, that which primarily distinguishes between a self-organizing and an artificial system is that in the former case the goal comes from within the system, that is the system designs itself and for itself (Arnellos, Spyrou, & Darzentas, 2007a), while in the latter case, the goal comes from an external designer. For the moment, we are not going to comment in the respective difference. What is really important is that a certain kind of functionality emerges out of the code's selection and interrelation processes and it emerges in such a way, that one is justified to say that it is the new functional organization that establishes the autonomy of the system (no. 2 in Varela's quotation), but it is also responsible for its maintenance, as it is the code which continuously selects and interrelates the emergent processes with a focus on the goal of their regeneration and realization (downward causation), that is, with a focus on its self-maintenance (no. 1 in Varela's quotation). It appears that for Varela, from an epistemological perspective, autonomy is equivalent to the notion of self-referentiality, which in turn, it is connected to the concept of organizational closure (Luisi, 2003). The basis of Varela's conception of autonomy is its active role in the contribution of the self-maintenance of the autopoietic system and especially in the production of its active components, but also in the effective alteration of its boundary

conditions in order for the system to be able to maintain its necessary variables in a homeostatic way.

On the same track, cognitive behavior is the result of a higher level of autonomy, where the neural system creates invariant patterns of sensorimotor correspondences in order to determine the behavior of the living system as a unit that exists and acts in space. Specifically, von Foerster, in his really enlightening attempt to describe the self-referential nature of the nervous system, he makes a model of it as a matrix (Fig. 2) in which the squares are nervous cells and in between them are the synaptic gaps, which can be filled with transmitters and says that:

> The black squares labelled N represent bundles of neurons that synapse with neurons of other bundles over the (synaptic) gaps indicated by the spaces between squares. The sensory surface (SS) of the organism is to the left, its motor surface (MS) to the right, and the neuropituitary (NP), the strongly innervated master gland that regulates the entire endocrinal system, is the stippled lower boundary of the array of squares. Nerve impulses travelling horizontally (from left to right) ultimately act on the motor surface (MS) whose changes (movements) are immediately sensed by the sensory surface (SS), as suggested by the "external" pathway following the arrows. Impulses travelling vertically (from top to bottom) stimulate the neuropituitary (NP), whose activity release steroids into the synaptic gaps, as suggested by the wiggly terminations of the lines following the arrow, and thus modify the modus operandi of all synaptic junctures, hence, the modus operandi of the system as a whole. Note the double closure of the system that now recursively operates not only on what it "sees", but on its operators as well. (von Foerster, 1988/2003, p. 225)

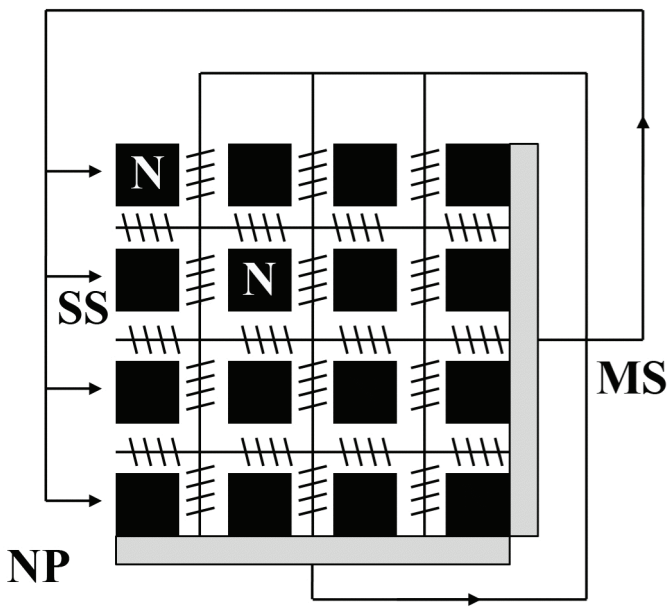



Fig. 2. Double closure of the senso-motoric and inner-secretoric-neuronal circuits

He then continues by arguing that this double closure between these two non-trivial machines are making computations which are subject to a non-trivial constraint, which is postulated as cognitive homeostasis, namely that "The nervous system is organized (or organizes itself) so that it computes a stable reality" (von Foerster, 1988/2003, p. 225).

Again, instances of emergence and downward causation are present in this organization, based on which, the autonomous system is able to perceive and act in its environment by internally creating meaningful information about its external environment. Specifically, a new functional meaning (modus operandi) emerges out of the self-organizing activity of the sensorimotor system with the activity of the neurohypophysis that produces steroids and releases them in the synaptic gaps. This emergent functionality forms the neural system as a whole, through which the agent interacts with the world and then immediately feeds back into it the respective environmental changes, which, through the receptors, return to the motor system, and in turn, regulate its self-organization (downward causation).

Another important issue that should be noticed is that in the second-order cybernetic framework a certain kind of autonomy is established, where the cognitive capacities are directly related with the capacity of the system to be alive. Particularly, in this perspective of agency, intentionality and especially, the endogenous production of purpose are located at the level of the origin of life and of biological functionality. Therefore, this inclination of a self-organizing cognitive system to maintain its own self-organization constitutes the core of its intentional and purposeful (goal-oriented) interaction with the environment. This is another characteristic of autonomous self-organizing and autopoietic agents, which distinguishes them from artificial agents (see section 3).

The analysis of the functional formation of the main concepts related to autonomous agents, presented so far, could also be used as a basis for judging the autonomy of artificial agents. However, the respective description is not adequately naturalized (Arnellos, Spyrou, & Darzentas, in press). Although we managed to ground autonomous agency on the functionality of the self-organizing system and to introduce some requirements for it, this is not as far as we may go in terms of naturalization and hence, this will not be the most appropriate theoretical ground based on which we will be able to judge for their autonomy, and most importantly, to advise their design. Therefore, we continue our analysis with the relevant, to our problem, notion of cohesion.

### *2.2 Emergence of Cohesion via Process Closure*

In section 2.1 it was argued that what defines an autonomous system is a global network of relations that establishes some self-maintained dynamics, where action and constitution are identical properties of the system itself. Practically, this means that the activity of the system is constituted of the constant regeneration of all the processes and of the components that constitute the system as an emergent functional whole. It is due to self-reference that the organizationally closed nature of such an

autonomous system is not considered as circular. Actually, the internal productive interrelations acquire a cohesive functional meaning in a collective way, since they contribute to the overall maintenance of the system. In the respective organization the whole and the parts are correlated to each other in a highly dynamic and reciprocal way. This systemic pattern of organizational (functional) dynamics is observed in every self-organizing system. Collier (1988) and Collier and Muller (1998) have called this pattern of organizational dynamics as cohesion, which is an inclusive capacity of an autonomous system and it indicates the existence of causal interactions among the components of the system in which certain capacities emerge and hence, the respective components are constituents for the system itself. Cohesion is not an epiphenomenal property, but on the contrary, it is exactly the emergence of this functional cohesion that avoids meaningless circularity and as such, the organization of the respective autonomous systems disregards the classical mechanistic opposition between the constituent parts and the global properties of the system itself.

Cohesion is an emergent property and as such, it can only be explained with respect to the causal roles that the constituent components and the relations among them acquire in the dynamic organization of the system.

Cohesive systems exhibit different kinds of correlations between different processes with respect to the degree (or the type) of cohesion that they exhibit. Systems with very strong and highly local bonds exhibit a powerful cohesion, which does not necessarily provide them with genuine autonomy and agency. Nevertheless, in the level of autopoiesis, or in what Ruiz-Mirazo & Moreno (2000) call the level of metabolic agency, the respective cohesion emerges in systems that are thermodynamically open and function in far-from-equilibrium conditions (Collier & Hooker, 1999). Such systems exhibit a kind of long-range correlations between different processes (certainly longer than the correlations that one can meet in a rock or in a self-organized crystal). As Collier (2007) has stressed, since there is an internal need for the coordination of the processes in order for them to achieve viability (self-maintenance), one should expect to find in such an autonomous system a holistic organization in which organizationally/operationally open aspects of lower level are closed at higher organizational levels. As it has already been argued, this is a highly constructive type of autonomy and it requires what Collier (1999) suggests as process closure (in accordance with organizational/operational closure), in order to mention the fact that in such autonomous systems there are some internal constraints controlling the internal flow of matter and energy, and by doing so, the whole system acquires the capacity to carry out the respective processes, since these processes will contribute to its self-maintenance.

Furthermore, and as it appears from Varela's quotation, the nature of the emergent process closure implies that all the interactive alternatives of the cognitive system are internally generated and their selection is an entirely internal process. Therefore, such autonomous systems must construct their reality by using internally available structures. Their functionality is entirely dependent on its structural components and their interrelationships that establish the respective dynamics. Hence, the functionality

of the cognitive system is immediately related to the maintenance of its systemic cohesion and consequently of its self-organizational dynamics. At this point, one should notice the interesting relation between second-order cybernetic systems, or systems that emerge functional cohesion mainly through process closure, with von Uexküll's theories about the functionality of living systems (von Uexküll, 1982). For von Uexküll, living organisms contain a functional rule or a building plan (i.e. an organizational code), which has an inherent meaning quality and thus, living systems are acting plans (i.e. they design themselves and for themselves, see also Arnellos, Spyrou, & Darzentas, 2007a), in contrast to machines that act according to the plans of their designers. This capacity makes them autonomous systems, such as the autopoietic systems, in contrast to machines, which are allopoietic systems.

What is really important regarding this type of emergent coherence is that since, in the more general second-order cybernetic framework of autopoiesis, functional closure enables the recursively interdependent generation and realization of the involved processes themselves, what really emerges is a distinct autonomous agent with a simultaneously configured world of perception and action. This is exactly what von Uexküll calls the coming together of the organism's components to form a coherent whole, which acts as a subject. This is the reason for arguing that this emergent coherence forms a system whose cognitive capacities are directly related with the capacity of the system to be alive. The emergent coherence is a result of a functional embedding, and it provides such autonomous systems with a certain kind of embodiment, which make the study of their behavior irreducible to physics and chemistry. For von Uexküll, this kind of embodiment emerges the Umwelt (i.e. the subjective world) of the autonomous agent.

In the autonomous systems described so far, perception and action are so closely related to the self-constructive and self-maintaining dynamics of the system. As a result, any downward causation, as the constraining of the function of the system's parts from the whole, will also acquire very fast and local characteristics in order to be able to be synchronized with the next step of the functional emergence, since the fundamental purpose of such systems is to maintain themselves. This is a kind of strong downward causation that comes in a greater degree from the higher levels of the system than from the environment.

Nevertheless, rocks and crystals show great degrees of cohesion with the respective emergent and downward constraining characteristics, but they are not exhibiting any significant intentionality, let alone experiencing any Umwelt, and as such, they cannot be considered as genuine agents. Agency do comes in a lot of degrees and different levels in nature (Arnellos, Spyrou, & Darzentas, in press), but almost everybody would agree that living systems are quite different from rocks. The main difference lies in the fact that genuine autonomous systems, such as living systems, exhibit a high degree of disentanglement from the environment, not in terms of their interactive processes, but, in terms of their ability to adapt in different environmental perturbations. On the contrary, systems merely exhibiting cohesion via process closure emerge a functional organization that is too tight with their

environments, but with minimal interactive characteristics, and as such, they cannot evolve beyond a certain threshold. Hence, such systems are at the threshold of autonomy exhibiting, at most, a reactive type of agency.

It will then be safe to argue that cohesion via process closure is a necessary but not a sufficient condition for genuine autonomy and agency. Again, there are enough tools to judge and criticize the autonomy of contemporary artificial agents, but there are also some other important issues that should be considered in order to better to advise their design. Particularly, genuinely autonomous agency is open-ended and emerges out of intentional and mostly ill-defined goals and purposes of the respective systems (Arnellos, Spyrou, & Darzentas, 2007b). Therefore, agency cannot be solely a matter of internal constructive processes and process closure. The need for open-endedness calls for interaction of the autonomous agent with the environment, while, the functional aspects of such an embodiment and its anticipatory content calls for advanced and efficient mechanisms of controlling and managing these interactions.

### *2.3 Emergence of Normative Functionality*

As it was described in 2.2, and due to the organization code – the functional rule, or the building plan of the system – a qualitative and quantitative imbalance emerges that indicates an asymmetry between the system and its environment. Specifically, in the self-organizing systems described so far, this asymmetry is created and maintained by the functionality of the system through the establishment of internal constructive relations that differentiate the system from its environment organizationally, and further, specify its autonomy and its identity. Hoffmeyer (1998) strongly argues that the secret of life and the development of agency are constructed upon this fundamental asymmetry. From a biological point of view, Hoffmeyer (1998), and others (see e.g. Ruiz-Mirazo & Moreno, 2000, 2004) suggest that this asymmetry is produced via a semi-permeable membrane. This membrane plays the role of dynamic boundaries, which has a functional basis of a chemical nature as they are the result of a productive organization and of the activity of the self-regulating and self-modifying processes of their systems.

This self-regulation aims in the maintenance of the system. The autopoietic model exemplifies this active relation between the boundary and the recursive production processes of the system's constitutive components, but with an emphasis in the absoluteness of the control and constrain of the flows of energy and matter in the system from the environment (Collier, 2004b). As also suggested by Hoffmeyer, this relation is a relation of regulation, hence, it cannot be an absolute one. Although the material basis of the complex boundary that supports the asymmetry is crucial for the functional emergence of such a boundary, for the moment, let's stay with the logical implications of such an asymmetry.

Bickhard (1993, 2000) exemplifies the implications of this asymmetry by postulating a recursive self-maintenant system, which is a self-organizing system that has more than one means at its disposal in order to maintain its ability of being self-maintenant in various environmental conditions. This is a self-organizing system

which functions far-from the thermodynamic equilibrium by continuously interacting with the environment, from where it finds the appropriate conditions for the success of its functional processes. Far from equilibrium processes cannot be kept in isolation, as they will run out of their dynamic functional stability. Consequently, the interactive opening of the system to the environment is considered as the most important point in its evolution towards genuine autonomy and agency, as it first of all enhances the stability of the system and its ability to maintain its maintenance. Specifically, the interactions in which an autonomous agent engages will be functional and dysfunctional (Moreno & Barandiaran, 2004). The former corresponds to the interactions which are integrated in the functional organization of the agent and in this way they contribute to its self-maintenance. The latter corresponds to the interactions that cannot be properly integrated in the functional organization and hence, they do not contribute or/and disturb the self-maintenance of the system.

Therefore, the primary goal of such a self-organizing system is to maintain its autonomy in the course of interactions. Since it is a self-organizing system, its embodiment is of a kind that its functionality is immediately related to its autonomy, through the fact that its apparent inclination to maintain its autonomy, in terms of its self-maintenance (its purpose), constitutes the intentionality of its actions and hence, of its interaction with the environment. As such, autonomous systems do not only exhibit process closure, but also interaction closure (Collier, 1999, 2000, 2007), a situation where the internal outcomes of the interactions of the autonomous system with its environment contributes to the maintenance of the functional (constructive/interactive) processes of the system that are responsible for these specific interactions. It is cohesion via process and interaction closure that distinguishes truly autonomous systems from other kind of cohesive systems. In this case, an autonomous system is not only able to maintain itself, but it can also meaningfully alter its internal functionality in order to adapt to complex and changing conditions around the environment. This capacity for meaningful critique regarding the functional and the dysfunctional with respect to the maintenance of the system is a normative one. Self-maintenant systems that exhibit normative functionality are truly autonomous systems and they present genuine agency (for more details on normativity and agency see Moreno & Barandiaran, 2004; Bickhard, 2005; Arnellos, Spyrou, & Darzentas, in press).

In this way, the overall functional closure (process and interaction closure) of an agent is guided by its autonomy, in the sense of the former contributing for the maintenance of the latter, while its intentionality derives from this specific normative functionality, as the latter is being directed towards the primary purpose of maintaining the self-maintenance. This cohesive combination of process and interaction closure is responsible for the emergence of functional norms within the autonomous system and for the autonomous system itself. Emmeche (2000), being on the same track, says that "the notion of function in biology is the teleological notion of 'a part existing for the good of the whole,' or 'having the purpose of' doing something in relation to the whole." (Emmeche, 2000, p. 194). As such, he also adopts the

normative perspective of emergent functionality, while he also suggests that functionality is only possible under a closure of operations, but as the capacity for interaction closure suggests, he also argues in favor of an only partial and relatively open functional closure. Specifically, Emmeche says that:

> Only when the causal chain from one part to the next closes or feeds back in a closed loop -- at once a feed-back on the level of parts and an emergent function defined (as mentioned) as a part-whole relation -- can we talk about a genuine function. In other words: It is because function is the function of a part that works effectively to produce (part-part efficient causation) influences on other parts within the same whole (the same form, the organism's) -- where each part is constrained by the same whole (formal causation) -- the total of parts interacting under these constraints in a coherent emergent pattern *is* the whole organism, whose maintenance (final causation) as form is the goal of each part. Here, final causation -- i.e., the dual process of *downward* constraints (formal cause) on the behavior of the parts and the *emergent* pattern of the parts forming a functioning organism (final causation), which is made of parts (material causation) -- is the causation of a physical part within a biological whole being committed to a specific role in the internal organization of that whole, thus the internal ascription (*de re*) of a role to the part is the emergence of that part's function. (Emmeche, 2000, p. 195)

What Emmeche tries to indicate is that a certain part or process of a system serves a function as far as it, first of all, contributes to the maintenance of this system and the role of this part or process is emergent in the internal organization of the respective whole, while this whole is downwardly constraining the emergent pattern/form of this part or process. Hence, normative functions emerge as a contribution for the autonomy of the agent, and with the goal of satisfying the respective functional norms.

What is still missing is meaning, on the basis of which the cognitive system decides which of the available functional processes should make use of, in order to successfully interact with a specific environment, that is, in order to fulfill its goal, that is, to satisfy its functional norms. In this case, an autonomous system uses its anticipations with the respective representational content (meaning). But, where exactly is this content to be found?

### *2.4 Anticipations and the Emergence of Representational Content*

Bickhard argues that such an autonomous system should have a way of differentiating the environments with which it interacts, and a switching mechanism to choose among the appropriate internal functional processes that it will use in the interaction. The differentiations are implicitly and interactively defined, as the internal outcomes of the interaction, which in turn depends on the functional organization of the participating subsystems and of the environment. These differentiations create an epistemic contact with the environment, but they do not carry, in any way, any representational content. However, they are indications of the interactive potentiality of the functional processes of the autonomous system itself. As such, these differentiations functionally indicate that some type of interaction is available in the specific environment and hence, implicitly predicate that the environment exhibits the appropriate conditions for the success of the indicated interaction.

In this model (Bickhard, 1993, 2000), such differentiated indications constitute emergent representations. The conditions of the environment that are functionally and implicitly predicated by the differentiation, as well as, the internal conditions of the autonomous cognitive system (i.e. other functional processes or conditions), that are supposed to be supporting the selected type of interaction, constitute the dynamic presuppositions of the functional processes that will guide the interaction. These presuppositions constitute the representational content of the autonomous cognitive system regarding the differentiated environment. This content emerges in the interaction of the system with the environment. What remains to be shown is how this representational content is related to the anticipations of an autonomous system.

Anticipation relates the present action of an agent with its future state. An anticipatory system has the ability to organize its functional state, in such a way that its current behavior will provide the ability to successfully interact with its environment in the future. An anticipatory system needs to be able to take into consideration the possible results of its actions in advance (that is, prior to its action and as such, purely reactive systems are not capable of anticipative functionality), hence, anticipation is immediately related to the meaning of the representations of the autonomous cognitive system (Collier, 1999). In this way, anticipation is one of the most characteristic aspects of autonomous systems due to their need to shape their dynamic interaction with the environment so as to achieve future outcomes (goals of the system) that will enhance their autonomy. In the context of the autonomous systems discussed so far, these future outcomes should satisfy the demand for process and interaction closure of the system and in general, for system's normative functionality.

Normative functionality is evaluated on the basis of the functional outcomes of the autonomous system, therefore, anticipation is immediately related to functionality (Collier, 2007). Even the simplest function requires anticipation in order to be effective. As mentioned before, anticipation is goal-directed. As a matter of fact, anticipation almost always requires functionality, which is, by default, a goal-oriented process. In this perspective, anticipation guides the functionality of the system through its representational content. In the model of the emergence of representations in the special case of an autonomous agent presented above, the representational content emerges in system's anticipation of interactive capabilities (Bickhard, 2001). In other words, the interactive capabilities are constituted as anticipation and it is this anticipation that could be inappropriate and this is detectable by the system itself, since such anticipation is embedded in the functional context of a goal-directed system (the emergent normativity).

These anticipations are guiding the interpretive interactions of an autonomous agent, that is, the recursive regulatory relations between itself and its environment. In case these interactions contribute to the agent's self-maintenance, its capability for interactive anticipation progressively increases and as such its intentional capacity increases too (Christensen & Hooker, 2002; Arnellos, Spyrou, & Darzentas, 2007b, in press).

So far, an analysis of the functional emergence of the fundamental properties of an autonomous agent has been provided. In the next section, this analysis will be used as a theoretical ground for the evaluation of the autonomy of contemporary artificial agents. The point of reference for this evaluation will be the symbol-grounding problem, which as it will be shown below, it is directly related to the emergence of functionality in an autonomous system.

## 3 How Autonomous is an Artificial Agent?

*3.1 The Symbol-Grounding Problem as an Implication of non-Emergent Functionality*

Almost every attempt to build an artificial agent begins by trying to connect the internal world of an agent with its external environment. Most times this connection is being made through the use of symbols, where each one of them has a meaning related to a state of affairs in the external environment. These symbols are playing the role of representations connected with the action modules of the artificial system (i.e. software, hardware or any degree of their combination). The processing of these symbols results in new meanings which guide the action of the system towards its environment. The disembodied nature of these symbolic systems results in the formulation of representations with no connection or/and correlation with the structure and the functionality of the system. This is the essence of the symbol-grounding problem, which comes as a set of problems posed by (Harnad, 1990), who founded his attack on Searle's Chinese Room Argument (Searle, 1980).

Harnad's argument was that an artificial agent does not have access to the meaning of the symbols it manipulates, but, the observer ascribes meaning in its actions. This is like somebody is trying to learn Chinese from a Chinese to Chinese dictionary. He will be able to reply to a Chinese question with a Chinese answer (provided that this is a super-efficient dictionary), but he will never be able to grasp the meaning of Chinese words. In other words, how can syntax ever acquire a semantic content? Therefore, based on the direct analogy, Harnad argues that computers and the respective agents will never be able to grasp the meaning of the symbols they manipulate, and as such, they will never be able to semantically connect these symbols with the respective state of affairs of the environment with which they interact. Artificial agents will never be able to develop the capacity for autonomy.

The source of the symbol-grounding problem is the grounding of meaning within the autonomous system itself. If we accept the analysis of section 2, then, intrinsic meaning requires intrinsic intentionality, which will provide the appropriate functionality for the emergence of meaning, that is, for the emergence of new types of functionality, which will result in new meanings, which will contribute to the autonomy of the system. Along the same lines, Collier (1999) suggests that the prerequisite for representational autonomy (which will immediately vanishes the symbol-grounding problem) is the emergence of functional autonomy from embodied intentionality. As it was argued in 2.4, the meaning of representations is directly related to the anticipations of the system. In case anticipations are not functional

emergents of the system, the latter will not be able to confront any environmental change beyond those for which it has been designed to. This is an artificial system with no inherent but with a derivative intentionality, and any such system functions in accordance to the anticipations of its designer, hence, it is design limited. Moreover, such artificial systems cannot alter or enhance their anticipations on their own in order to achieve greater flexibility for their interaction with the environment.

As it has already been shown, these problems will prevent anyone from calling the respective artificial systems as autonomous agents. However, as it was mentioned in the beginning of section 2, almost all the design results of the disciplines of the new AI and of robotics are called as autonomous agents. This is, primarily, because most of the researchers consider symbol-grounding as a problem concerning only the computational framework of cognition (see e.g., Fodor, 1990) and its cornerstone, the physical symbol systems hypothesis (Newell, 1980). Additionally, there is a huge amount of research trying to analyze or/and solve the symbol-grounding problem (see e.g., Chalmers, 1992; Ziemke, 1999; Coradeschi & Saffioti, 2003), and a lot of researchers arguing in favor of it having been solved (for a different kind of analysis regarding the efficiency of several proposed architectures for solving the symbol-grounding problem, see Taddeo & Floridi, 2005). In the next sections we will examine the main approaches in the solution of the symbol-grounding problem having as a basis the theoretical analysis of the functional emergence of a genuine autonomous agent that was presented in section 2.

### *3.2 Emergence in Computational/Representational Agents*

#### 3.2.1 Computationalism provides emergent correspondences

As a solution to the symbol-grounding problem, Harnad suggested a hybrid symbolic/connectionistic system where symbolic representations are grounded in two types of non-symbolic representations: a. in the iconic representations, which are analog transformations of sensorial perceptions and b. in categorical representations, which take advantage of the sensorimotor invariants for the active transduction of sensorial perceptions to basic symbols (e.g. horse, stripes), from which more composite symbolic representations can be build (e.g. zebra = horse + stripes). In other words, categorical representations are the elements of a systematically combinatorial system. Harnad proposes the use of neural networks for the bottom-up transformation of the real world's objects to individual symbolic representations through the use of non-symbolic representations (Harnad, 1990, 1993).

Harnad argues that the respective categorical representations result from keeping only the invariant properties of an iconic representation, so the cognitive agent will be able to recognize and not only to discriminate an object. He of course admits that this is very difficult to be implemented as it involves the physiology of perception of the natural cognitive systems. Another difficulty, which is also recognized by Harnad, is that in his approach (and in almost all approaches in building an artificial agent) there are some logical operators that should be externally imposed (by the designer) in order

the system that combines the categorical representations to function and to configure the respective symbolisms.

The problem is concentrated on the way that the meaning of these operators will be imposed to the system. It is obvious that this kind of meaning does not emerge within the system and the same goes for the respective functionality which is driven by this very meaning. Harnad (1995) argues that such architecture needs a robotic functionality and not a merely computational system, in order for the categorical invariants to be grounded in a realistic (and not in a virtual) sensorimotor interaction between the system and its environment.

What is missing from Harnad's solution regarding the symbol-grounding problem, and, indeed, from every solution which is provided under the cognitivistic/connectionistic framework, is not the way that the categorical representations are formed, but the need for a clarification of the relation between the external signal and of its iconic representation as its analogue. If this is to be made via a simple transduction of the respective signal, then, the respective correspondence would count as a representation. The problem at this point is that one cannot analyze this transduction and hence, nobody knows if such a correspondence could really count as a representation. Indeed, if one considers the framework of second-order cybernetics described in section 2.1, he will conclude that there is no space for any kind of direct or indirect correspondences, and furthermore, one is not justified to say that *aboutness* comes as a function of such correspondences.

On the other hand, one should also try to explain how such correspondence pre-exists in a respective representation, which is equivalent to explain the whole physiology of perception. Harnad and all other similar approaches use the notion of information as a magical quantity which exists in the external signal and somehow is passed in the representation. This would be acceptable if we had such a theory. For the moment we do not, and considering the approaches presented in section 2, it is highly likely that information is not something that can be passed from one cognitive agent to another, but it is rather the result of the functional formation of a system, and as such, it belongs to it and it stays within it. This is highly related and of course, in accordance with Bickhard's (1993) suggestion, namely that emergent representation, that is, emergent meaning needs inherent aims and goals, otherwise, all one may have is correspondences with no grounded meaning in the artificial agent, but grounded only in the mind of its designer.

### 3.2.2 Computational and Weak Emergence

There are numerous approaches under the cognitivist/representationalist umbrella, which are all facing the same fundamental problem: what emerge are correspondences and not representations because of the derivative nature of the respective functionality, and as such, there is no emergent meaning, hence, there is no genuine autonomy. Such examples are the approaches of (Cangelosi, Greco, & Harnad, 2000) and (Cangelosi & Harnad, 2001), where they use a very complicated three-layered feedforward neural network as the transformer of categorical perception into grounded low-level labels

and then, into higher-level symbols. Although the main problem, that is, the transformation of external data to semantic content for the machine, remains, there are some other issues that should be taken under consideration.

In a feedforward neural network activation is propagated in only one direction (from input to output) and after some time, where the network will have been trained and the weights of its connections will have been stabilized, the respective mappings will remain the same, as the network cannot alter its transfer function. In this case, as Ziemke and Sharkey (2001) argue, the network becomes a trivial machine (to use von Foerster's words), or in other words, a passive action-reaction system. Therefore, these kinds of architectures are breaking down to systems that solely emerge correspondences and not in an open-ended way, as after some time the mapping remains invariant.

However, for their authors, the merit of these architectures is not only the very powerful transduction mechanism, but also the combinatorial strength of the learning modules that are fed by the transducer and they can provide the artificial agent with a very rich vocabulary of higher-order concepts and of language. The appearance of a variety of high-level concepts in these systems is considered as a case of genuine emergent behavior, and the respective systems are considered as autonomous. The reasons for the lack of some concepts in these systems is that the selected underlying functionality or/and the learning mechanism is not the best possible, or that they system has not still interacted with the variety of the environments that a natural agent needs to interact in order to emerge a great variety of meaning.

The acceptance of this kind of emergence, either in concepts or in primitive behaviors is also evident in the domain of ALife Representative paradigms are the ones of Langton (1989) and of Baas (1994). Langton argues in favor of the emergence of genuine life in artificial systems as the result of a mapping of the low-level behaviors of the simulated natural systems (e.g., bird flocking) into informational computer processes. The emergence of a higher-level behavior is not only simulated, but it realizes the same thing with the natural phenomenon. Baas proposes complexity, hierarchies emergence and evolution as four interrelated phenomena which every biological system presents and which are also supported in ALife simulations. He also suggests that in ALife simulations one can observe both emergence and downward causation, and he also argues that these two properties can only empirically be proved that are being exhibited by a natural system.

The theoretical justification of these claims comes from the work of Bedau (1997; 2002), who proposes that in such computational simulations there is the appearance of weak emergence as there are new emergent macroscopic states that can be derived from the microscopic dynamics but only through simulations. This kind of emergence is in principle predictable, but not in every detail, as the weakly emergent properties arise from the top-down feedback processes (downward causation). In weak emergence there is no unique direction of causality from the microscopic to the macroscopic level (as for example in feedforward neural networks), but there are

causal relations in both directions. It is probably of no need to say that this kind of approach to the emergence of new behavior in artificial systems is dominant in ALife.

However, the problems are numerous and this kind of emergence has been attacked from many thinkers (see e.g., Cariani, 1991; Kampis, 1991; Emmeche, 1992, 1994) with respect to its relation to genuine emergence and the consequent autonomy. The conclusion of these critiques is that this is a computational emergence in which global patterns arise from local micro-deterministic computational interaction. Any finite-state machine which is used in these computations is a determined machine with predetermined transition rules and predefined primitives. As such, it is like somebody trying to simulate via prior selected rules the interpretations of a specific process, while this very interpretation has already been realized by an external natural cognitive system (the designer of the simulation). Additionally, such kinds of simulations are by default disembodied, while as argued in section 2, autonomy requires a body and a respective embeddedness in the environment from which it emerges. As Kampis (1991), Emmeche (1992), have suggested, formal computation does not have the causality of natural causation, or as von Neumman (1966) argues, by adopting only the logical part of a process (its abstraction) we may lose the most interesting part, that is, its material basis and the respective causality.

So, computational emergence is in no way a genuine functional emergence, and on a basic level, it is not significantly different from any other kind of computationalism. The interrelations between intentionality, functionality and meaning do not hold, or to be more specific, they do not even exist, hence, these kinds of systems cannot be considered as autonomous systems.

### *3.3 Emergence in Physically Grounded Artificial Agents*

#### 3.3.1 Emergence in the subsumption architecture

As it was mentioned in the previous section, one of the main points in Harnad's approach to the solution of the symbol-grounding problem was that symbol-grounding is an empirical issue and then, one needs a robotic functionality for the perceptive invariants to be grounded in a realistic and not in a virtual (simulated in software) sensorimotor interaction. In this way, the agent will be physically grounded, hence, it will be situated and embodied. The first and pioneered attempt toward this direction came from Brooks (1986, 1993), who introduced the subsumption architecture. There is no need to analyze the specific architecture, as it is well-known for its merits and for its disadvantages (see e.g., Emmeche, 2001; Christensen & Hooker, 2004), but some things relevant to its allegedly emergent functionality should be put under consideration.

The main concern of Brooks was to design an agent who would be able to interact with its environment in real-time, so that it will be able to confront real-life situations. The dominant computational approach, strongly influenced by the computer-based metaphor of the mind, requires that an agent will first sense its environment, it will then think and at the end it will act. This approach demands an a-priori determined and imposed representational model of the world, which will guide the central

processing unit of the system. Such an architecture cannot cope with the enormous variety of a real-life environment, hence, something else is needed.

Brooks proposed the subsumption architecture as an alternative to the representational/computational model of the mind. The subsumption architecture has no central controller, but on the contrary, global control emerges out of the interaction of hierarchically organized behavioral units of the system. For example, the control of a simple robot wandering around a room trying to avoid certain obstacles emerges out of one behavioral unit that makes the robot to move forward and from a second unit, which, every time the robot meets an obstacle, subsumes the first unit and makes the robot to turn towards another direction. The subsumption architecture begins with simple functional units supporting fundamental activities of the agent, the interactive capacity of which increases with the addition of other more elaborated levels of action. Such an artificial agent presents the following characteristics: distributed control, direct coupling between perception and action, cohesion between multiple hierarchically organized functional modules, interaction based on its own functionality and not through some abstract and ungrounded representations, action through the maintenance of the functional cohesion and taking under consideration its aims and goals, dynamic interaction with the environment and of course, situatedness and embodiment (physical grounding).

Considering the list with the characteristics of such an artificial agent in comparison with the properties of an autonomous agent sketched in section 2, one could assume that this is a genuine autonomous agent. Of course, this is not the case. An agent with the subsumption architecture exhibits no central control and this is something that reminds us of the self-organizing and autopoietic systems, where all the functional processes of the system are responsible for the emergence of novel organizations, hence, of emergent functionality. Additionally, such artificial systems make no use of representations, since their interaction is directly guided by the functionality of the respective modules engaging in the interaction. This is also something that pertains to the characteristics of second-order cybernetic systems, which ascribe the existence of representation in the eye of the observer. Therefore, such an artificial agent presents a direct coupling between perception and action, which results in a kind of weak structural coupling with its environment. This directness, which comes as a result of the agent's physical grounding, practically vanishes the symbol-grounding problem. The agent uses its functionally integrated meaning to guide its interaction and this guidance supports the cohesion of its functional levels in accordance with its goal. Based on the analysis of 2.2 it could be said that these characteristics are just the results of an emergent functionality via process closure, which results in a cohesion maintained by the interrelations of the functional modules. The closure of the process of each module will either be satisfied (will not loop forever and it will pass execution on another module) in the same module or in another module in case a subsumed module takes control. In this perspective, the most important module of the system can be considered as the initiator of a downward causation which also propagates through several lower levels.

So far so good for the subsumption architecture, but there are no more good news. The exhibited cohesion is not genuinely emergent as most of the respective functionality is the result of an external design. Even if someone leaves aside this "small" detail, the respective cohesion exhibits very strong and local bonds, which do not provide the possibility for a great variety of actions. This is apparent in such an autonomous agent who resembles mostly an action-reaction system (as almost all systems that cannot surpass the level of autopoiesis or of metabolic autonomy). Additionally, and due to the rigidness of its cohesion, as well as due to the absoluteness of its boundaries (i.e. the rule-based and automata-driven input-output units of each module) such an artificial agent cannot scale on its own, unless new functional modules are added. Of course, one should not leave aside the fact that the building plan/organizational code of an agent designed based on the subsumption architecture belongs to its designer and not to the agent itself. Therefore, the goal of the system under which this functional cohesion is maintained does not belong to the system. Considering that derivative intentionality results in derivative meaning, the symbol-grounding problem is not solved, rather it is postponed until the time when the designer will decide about the functional modules of the system based on his anticipations.

### 3.3.2 Emergence in Agents With Artificial Nervous Systems

Researchers in autonomous and cognitive robotics and in adaptive systems have, in a way, tried to achieve a greater functional flexibility than the one presented in the rigid functional cohesion of the subsumption architecture by designing self-organizing artificial nervous systems. A prominent work is the one of Ziemke and Thieme (2002), where they use multiple recurrent neural networks (RNNs) with a second-order feedback, in order to simulate the sensorimotor system of a simple robot. Specifically, what they are trying to model is von Foerster's notion of double closure between the senso-motoric and inner-secretoric-neuronal circuits (see Fig. 2). The authors argue that with the suggested architecture the sensorimotor mapping changes dynamically with the internal state of the agent. In other words, the artificial nervous system changes its modus operandi. Ziemke (2005) has already make the connection between their design suggestion and von Foerster's double closure, but let's take a closer look to the allegedly emergent characteristics of such an artificial system.

What Ziemke and Thieme have tried to do is to design a cohesion closer to the one which emerges in genuine self-organizing systems. In particular, by functionally implementing a second-order feedback between the RNNs they acquire greater flexibility and a greater variety of interrelations between different organizational levels (i.e. input-output and the hidden units representing context). In this way various interrelations between different time scales can take place, driving the system's action in a somehow, non-derivative way. The closure conditions achieved through this setting resembles the closure conditions that a natural nervous system exhibits. For this to be done, the logical structure of the respective artificial nervous system is analogous to the functionally interrelated structure between the sensorimotor system

and the inner-secretoric-neuronal system. One has to admit that this is a type of cohesion which is closer to the one of natural self-organizing systems than the cohesion achieved in the subsumption architecture. Actually, Ziemke argues that the respective architecture result in the self-organization of the sensorimotor system of the robot. The interesting part would be to see if this new cohesion, which results from this kind of self-organization, offers any genuine emergence or any emergence at all.

First of all, it cannot be asserted that the respective artificial nervous system truly exhibit genuine self-organization. Collier, (2004a, p. 162) suggests up to six important characteristics of self-organizing systems, and mostly all of them are energetic characteristics, such as exportation of entropy, minimization of local entropy production, maximization of the efficiency of energy throughput under force, free energy source, phase separation and promotion of microscopic fluctuations to macroscopic order. It seems that what happens to the artificial nervous system which is simulated by RNNs is closer to re-organization that to self-organization. Practically, this means that the artificial nervous system does not emerge functional norms (see section 2.3), and how could have done this since it has been evolved independently of the robot's body. Hence, the nervous system alters its organization (although mostly in a resetting and recombining mode), but in which purpose and for the benefit of who? Certainly not for itself, because, first of all, the respective functionality (i.e. the selection of the RNNs and their functional interrelation, at least in the dimension of different levels), has been externally imposed. Secondly, and equally interestingly, because, even though the respective artificial nervous system exhibits an interesting process closure, which results in a certainly interesting cohesion, it cannot achieve the required interaction closure with its environment in an open-ended way. Hence, based on the analysis of section 2.3 and 2.4, this system cannot emerge functional norms and as such, it cannot emerge genuine representations. Although its coherence is an interesting one, its functional support results in the lack of normativity in the artificial agent. In other words, closure is necessary for functional emergence, but the endogenous evolution of closure is necessary for normative emergence and for the emergence of meaning for the system itself.

Normativity is a crucial issue for understanding the meaning processes in autonomous agents but it is highly neglected by the community of artificial systems research. Functional norms, in a way, attribute values of true or false, and they are emergent in system's interactions with the environment. Emmeche, rightly points out that a perceived sign may be the carrier of some general type, as danger, "but it has always also an aspect of being a tone, that is being qualitatively felt in some way (e.g., unpleasant)" (Emmeche, 2001, p. 680). The acceptance/understanding of such an unpleasant feeling and its consequent interpretation is probably the result of a normative functionality of an autonomous agent and of the respective anticipations with their representational content. For the moment, this is something that cannot be exhibited in silicon-based systems due to their dyadic nature, which first of all does not permit the emergence of complex and dynamically interactive boundaries, between the different functional levels of an artificial system and between the system

and its environment. Normativity is indirectly related to downward causation in an autonomous agent. Process and interaction closure in a self-maintained cohesion may require that low-level open issues will achieve closure in higher levels. On the other hand, the higher the autonomy and agency of a system, the higher the degree of abstraction of the concepts/meanings to which some of its norms can be related. In this case, the system should interact with the environment based on its anticipations, but for closure to be achieved, the emergent organizational level should functionally determine the lower level associated with the respective norm. It is obvious that this cannot happen with any self-organizing artificial nervous system which cannot at least exhibit the kind of cohesion that will functionally provide the conditions for the emergence of process and interaction closure. In this perspective, a robot with such an artificial nervous system is not much more than a rock with wheels.

#### 3.3.3 Emergence in *Self-organizing* and *Evolutionary* Artificial Agents

There is a great deal of research in the design and development of the so-called self-organizing agents that achieve an evolutionary adaptation with their environments. These attempts are characterized by self-organizing and evolution of robots bodies and controllers (see e.g., Nolfi & Floreano, 2000), and/or the so called self-organizing communication and evolution of languages (see e.g., Vogt, 2005). But again, all these works are falling under the same problems as those described in the previous section. Self-organization needs a self to organize (Collier, 2004b) and in all these cases, there is no self, at all. Imposed functionality in the form of an artificial ontogeny cannot create a genuinely autonomous system, as any kind of artificial ontogeny will result in the imposition of new functional norms, but not in their genuine emergence. As such, intentionality is still residing in the eye of the designer, or of the beholder. Imposed ontogeny which is not properly emerged cannot be functionally integrated with the building plan or the organizational code of the autonomous system, and as such, any structural emergence does not normatively serve the self-maintenance of the agent itself, but satisfies the emergent meaning that the designer himself associates with the aims that he has selected for its artificial agent.

The same goes for the allegedly emergent language and vocabularies in interacting robots. As it has been thoroughly analyzed in (Arnellos, Spyrou, & Darzentas, 2007b) communication between autonomous systems is crucial in order for an autonomous system to enhance its autonomy and consequently its cognitive capacities. For this to be done, the existence of inherent aims and goals, with a variable degree of definiteness, seems imperative, otherwise, the respective interactions will not have the respective emergent functional value, that is, new structures will be produced with no inherent grounding. Of course, ill-defined aims and goals, that is, higher-level anticipation emerges on the basis of endogenous and well-defined functional norms, which are grounded in the agent's self-maintenance. Contemporary artificial agents do not seem to exhibit such dispositions. At the end, all contemporary artificial agents are victims of their functionality in their attempt to overcome the symbol-grounding problem and to emerge new functional meaning.

## 4 Conclusions: Designing Representational Autonomous Agents

The critical review in section 4 is sure not an exhaustive one, but it is quite representative of the abilities of contemporary artificial agents and of their capacity for autonomous agency. The naturalistically emergent nature of agency (see Arnellos, Spyrou, & Darzentas, in press) does not allow for the partitioning of agency in *simpler problems* or the study of isolated cases of cognitive activity. Nevertheless, these phenomena are quite typical in the research of autonomous artificial agents. However, the notion of a simpler problem is always interpreted with respect to the theoretical framework upon which the design of the artificial agent relies.

In the more general domain of self-organization, where a systemically emergent perspective is adopted regarding the evolution of autonomous agency, the primary aim of an attempt to design an artificial autonomous agent is not to design an agent that will mimic in a great detail the activities of a human. Considering the analysis of section 2, this will probably demand the from scratch design and development of the extremely complex processes of life and of cognition combined with the evolution and the adaptation of the artificial agent. On the contrary, the aims of such research attempts should be the design of a complete artificial agent, that is, a design which will support, up to a certain satisfying level, the set of the fundamental and characteristic properties of autonomy, by maintaining its systemic and emergent nature in different types of dynamically changing environments. In this way, it is most probable that the design and development of an artificial agent that will tend to genuinely exhibit the emergent and interrelated properties of autonomy, intentionality, functionality and meaning will take a long time, but the respective trip will provide many interesting answers in a variety of really hard questions regarding the nature of an autonomous agent, while simultaneously will feedback and support the respective theoretical frameworks and models. The work of Ziemke (2005) is a work towards this direction, but it is still very difficult for the community of researchers to keep with the theoretical complexity and rigidity of the respective naturalized frameworks.

Considering the analysis of section 2, one may conclude that the design of an artificial autonomous agent requires the design of genuinely emergent representational autonomy. Such a system should emerge a functional cohesion which will allow for process and interaction closure. Process and interaction closure will provide the agent with the respective openness so that it will be able to follow the whole interactive cycle of its anticipations. This will result in a continuous emergence of new functional norms, and in consequence, of new meanings. The emergence of the proper type of cohesion cannot take place in an agent with a sensorimotor system, no matter its artificial variety and perplexity, which is being functionally separated, in terms of its structural evolution, with the body that it supposes to activate. Additionally, this emergence should be guided by the functional norms that it produces, namely, the functional requirement that every interaction of the system with its environment will be evaluated on the basis of its self-maintenance. This is a process of emergent functionality and of emergent meaning, which creates new functional organizations,

which in turn are downwardly constraining or determining the respective functionality.

What should be made clear is that simple logical or/and formal co-evolution of body and mind is not enough, as it seems that the energetic characteristics of self-organization proper are crucial for the emergence of the required functional normativity. Moreno and Etxeberria (2005) are right to argue that one cannot leave the energetic aspects aside and solely try to build sensorimotor autonomy instead of basic autonomy. As they suggest, the problem is not that computer power is still not enough, or that the mathematics should be reformulated, but that "The difficulty is in the deep and radical interrelation between forms of organization and materiality" (Moreno & Etxeberria, 2005, p. 173).

The need for different materiality should not come as a surprise, as it is something that other researchers, either intuitively (e.g. von Neuman, 1966), or quite thoroughly (e.g. Emmeche, 1992, 2001), Collier (1999, 2007) have also suggested. Indeed, to design an autonomous artificial agent is to design an artificial agent that is able to engage in design processes for itself and this is a genuine semiotic phenomenon (Arnellos, Spyrou, & Darzentas, 2007a), which demands the coevolution of the autonomous system with the mediators of the signs with which it interacts (Arnellos, Spyrou, & Darzentas, 2006). It is highly likely that silicon-based systems cannot support this requirement, but a carbon-based biology is needed. In any other case, the emergence of intentional behavior seems, for the moment, really impossible. However, a symbolically grounded agent is a genuinely emergent autonomous agent.

## References


Arnellos, A., Spyrou, T., & Darzentas, J. (2007a). Exploring creativity in the design process: A systems-semiotic perspective. *Cybernetics & Human Knowing, 14* (1), 37-64.

Arnellos, A., Spyrou, T., & Darzentas, J. (2007b). Cybernetic Embodiment and the role of autonomy in the design process. *Kybernetes, 36* (9-10), 1207-1224.

Arnellos, A., Spyrou, T., & Darzentas, J. (2006). Dynamic interactions in artificial environments: Causal and non-causal aspects for the emergence of meaning. *Journal of Systemics, Cybernetics and Informatics*, 3 (1), 82-89.

Arnellos, A., Spyrou, T., & Darzentas, J.(in press). Towards the Naturalization of Agency based on an Interactivist Account of Autonomy. *New Ideas in Psychology.* (special issue on Interactivism)

Baas, N. A. (1994). Hyperstructures – a framework for emergents, hierarchies, evolution and complexity. In C. G. Langton (Ed.), *Artificial Life III, Santa Fe Studies in the Sciences of Complexity, Proc. Volume XVII* (pp. 515-537). Redwood City, CA: Addison-Wesley.

Bedau, M.A. (1997). Weak emergence. *Philosophical Perspectives, 11,* 375-399

Bedau, M.A. (2002). Downward causation and the autonomy of weak emergence. Special issue on "Emergences and downward causation," *Principia, 6*, 5-50,

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence, 5*, 285-333.

Bickhard, M. H. (2000). Autonomy, function, and representation. *Communication and Cognition – Artificial Intelligence, 17*, (3-4), 111-131.

Bickhard, M. H. (2005). Consciousness and Reflective Consciousness. *Philosophical Psychology, 18*(2), 205-218.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation, 2*,14-23.

Brooks, R. A. (1993). The engineering of physical grounding. *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society* (pp. 153-154). Hillsdale, NJ: Lawrence Erlbaum.

Cangelosi, A., Greco, A., & Harnad, S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science, 12*, 143-162.

Cangelosi, A., & Harnad, S. (2000). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language perceptual categories. Special issue on "Grounding Language." *Evolution of Communication*, *4,* 117-142.

Cariani, P. (1991). Emergence and artificial life. In C. G. Langton, C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.), *Artificial Life II. Santa Fe Institute Studies in the Sciences of Complexity Proceedings Vol. X* (pp. 775-797). Redwood City, CA: Addison-Wesley.

Chalmers, D. (1992). Subsymbolic computation and the chinese room. In J. Dinsmore (Ed.), *The symbolic and connectionist paradigms: Closing the gap* (pp. 25-48). New York: Lawrence Erlbaum.

Christensen, W. D. & Hooker, C. A. (2002). Self-directed agents. In J. MacIntosh (Ed.), "Naturalism, evolution & intentionality," *Canadian Journal of Philosophy, Special Supplementary, 27*, 19-52.

Christensen, W. D., & Hooker, C. A. (2004). Representation and the meaning of life. In H. Clapin, P. Staines, P. Slezak, (Eds.), Representation in mind: New approaches to mental representation (pp. 41-70). Oxford: Elsevier.

Collier, J. (1988). Supervenience and reduction in biological hierarchies. In M. Matthen & B. Linsky (Eds.), "Philosophy and biology," *Canadian Journal of Philosophy Supplementary, 14* , 209-234.

Collier, J. (1999). Autonomy in anticipatory systems: Significance for functionality, intentionality and meaning. In D. M. Dubois (Ed.), *Computing Anticipatory Systems, CASYS'98 - Second International Conference, American Institute of Physics* (pp. 75-81). New York: Woodbury.

Collier, J. (2000). Autonomy and process closure as the basis for functionality. In J. L.R. Chandler and G. van de Vijver (Eds.), Closure: Emergent organizations and their dynamics. *The Annals of the New York Academy of Science, 901*, 280-291.

Collier, J. (2002). What is autonomy? *International Journal of Computing Anticipatory Systems*, *12*, 212-221. (Partial Proceedings of CASYS'01: Fifth International Conference on Computing Anticipatory Systems)

Collier, J. (2004a). Fundamental properties of self-organization. In V. Arshinov & C. Fuchs (Eds.), *Causality, emergence, self-organisation* (pp. 150-166). Moscow: NIA-Piroda.

Collier, J. (2004b). Self-organisation, individuation and identity. *Revue Internationale de Philosophie, 59*, 151-172.

Collier, J. (2007). Simulating autonomous anticipation: The importance of Dubois' conjecture. *BioSystems, 91* (2) 346-354.

Collier, J., & Hooker C. A. (1999). Complexly organised dynamical systems. Open systems and information *Dynamics, 6*, 241-302.

Collier, J., & Muller, S. (1998). The dynamical basis of emergence in natural hierarchies. In G. Farre and T. Oksala, (Eds.), Emergence, Complexity, Hierarchy and Organization: ECHO III Conference, Acta Polytechnica Scandinavica, MA91 Espoo: Finnish Academy of Technology.

Coradeschi S.& Saffioti, A. (2003). An introduction to the anchoring problem. *Robotics and Autonomous Systems, 43,* 85-96.

Emmeche, C. (1992). Modeling life: A note on the semiotics of emergence and computation in artificial and natural systems. In T. A. Sebeok & J. Umiker-Sebeok, (Eds.), *Biosemiotics: The semiotic web 1991* (pp. 77-99). Berlin: Mouton de Gruyter Publishers.

Emmeche, C. (1994). *The garden in the machine: The emerging science of artificial life*. Princeton, NJ: Princeton University Press.

Emmeche, C. (2000). Closure, function, emergence, semiosis and life: The same idea? Reflections on the concrete and the abstract in theoretical biology. pp. 187-197 In J. L. R. Chandler & G. Van de Vijver (Eds.), *Closure: Emergent organizations and their dynamics*. *Annals of the New York Academy of Sciences, Volume 901* (pp. 187-197). New York: The New York Academy of Sciences.

Emmeche, C. (2001. Does a robot have an Umwelt? Reflections on the qualitative biosemiotics of Jakob von Uexküll. *Semiotica, 134* (1/4), 653-693.

Fodor, J. A. (1990). *A theory of content and other essays*. Cambridge, MA: The MIT Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D 42*,335-346.

Harnad, S. (1993). Grounding symbols in the analog world with neural nets. *Think 2*(1) 12-78 (Special Issue on "Connectionism versus Symbolism" D.M.W. Powers & P. A. Flach, Eds.)

Harnad, S. (1995). Grounding symbolic capacity in robotic capacity. In L. Steels, & R. Brooks (Eds.), *The "artificial life" route to "artificial intelligence." Building situated embodied agents* (pp. 276-286). New Haven, CT: Lawrence Erlbaum.

Hexmoor, H., Castelfranchi, C., & Falcone, R. (Eds.) (2003). *Agent autonomy.* Boston: Kluwer Academic Publishers, Boston.

Hoffmeyer, J. (1998). Surfaces inside surfaces: On the origin of agency and life. *Cybernetics & Human Knowing, 5* (1), 33-42.

Kampis, G. (1991). *Self-modifying systems in biology and cognitive science*. Oxford: Pergamon Press.

Kampis, G. (1999). The natural history of agents. In L. Gulyás, G. Tatai, and J. Váncza (Eds.), *Agents everywhere* (pp. 24-48). Budapest: Springer.

Langton, C. G. (1989). Artificial Life. In C. G. Langton (Ed.), *Artificial life: Proceedings of the Santa Fe Institute Studies in the Sciences of Complexity,VI*, 1-47. Redwood City, CA.: Addison-Wesley.

Luisi, P. L. (2003). Autopoiesis: A review and a reappraisal. *Naturwissenschaften, 90*, 49-59.

Moreno, A., & Barandiaran, X. (2004). A naturalized account of the inside-outside dichotomy. *Philosophica, 73*, 11-26.

Moreno, A., & Etxeberria, A. (2005). Agency in natural and artificial systems. Almeida e Costa, F., Rocha, L. and Bedau, M. (Eds.) Special Issue on new robotics evolution and embodied cognition. *Artificial Life, 11* (1-2), 161-176.

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135-183.

Nolfi, S. & Floreano, D. (2000). Evolutionary robotics: The biology, intelligence, and technology of self-organizing machines. Cambridge, MA: The MIT Press.

Rocha, L. M. (1996). Eigenbehavior and symbols. *Systems Research, 13* (3), 371-384.

Ruiz-Mirazo, K., & Moreno, A. (2000). Searching for the roots of autonomy: The natural and artificial paradigms revisited. *Communication and Cognition – Artificial Intelligence, 17,* (3-4), 209-228.

Ruiz-Mirazo, K., & Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life, 10*, 235-259.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417-424.

Taddeo, M. & Floridi, L. (2005). The symbol grounding problem: A critical review of fifteen years of research. *Journal of Experimental and Theoretical Artificial Intelligence, 17*(4), 419 - 445.

Varela, F. (1979). *Principles of biological autonomy.* New York: Elsevier.

Varela, F., & Bourgine, P. (1992). Introduction: Towards a practice of autonomous systems. In F. Varela & P. Bourgine (Eds.), *Towards a practice of autonomous systems. Proceedings of the first european conference on artificial life*, (pp. xi-xvi). Cambridge, MA: The MIT Press.

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence, 167* (1-2), 206-242.

von Foerster, H. (1960/2003). On self-organizing systems and their environments. In *Understanding understanding: Essays on cybernetics and cognition* (pp. 1-19). New York: Springer-Verlag. (Originally published in M.C. Yvotis & S. Cameron (Eds.), Self-organizing Systems [pp. 31-50]. London: Pergamon, 1960)

von Foerster, H. (1976/2003). Objects: tokens for (eigen-) behaviors. In *Understanding understanding: Essays on cybernetics and cognition* (pp. 261-271). New York: Springer-Verlag. (originally published in *ASC Cybernetics Forum, 8* , 91-96, 1976)

von Foerster, H. (1981). *Observing systems*. Seaside, CA: Intersystems Publications.

von Foerster, H. (1988/2003). On Constructing a Reality. In *Understanding understanding: Essays on cybernetics and cognition* (pp. 211-228). New York: Springer-Verlag. (Originally published in *Adolescent Psychiatry, Developmental and Clinical Studies* [Vol. 15, pp. 77-95]. The University of Chicago Press, 1988)

von Neumann, J. (1966). *The theory of self-reproducing automata*. (A. W. Burks, Ed.). Urbana, IL: University of Illinois Press.

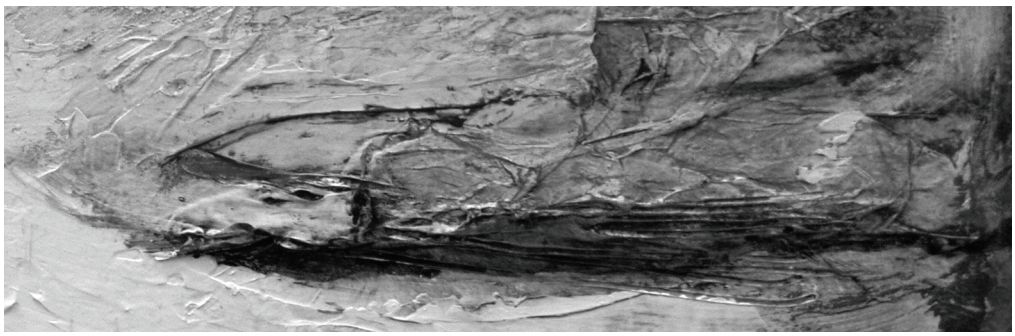
von Uexküll, J. (1982). The theory of meaning. *Semiotica, 42*(1):25-82.

Ziemke, T. (1999). Rethinking grounding. In A. Riegler, M. Peschl, & A von Stein (Eds.) *Understanding representation in the cognitive sciences* (pp. 177-190). New York: Plenum Press.

Ziemke, T. (2005). Cybernetics and embodied cognition: On the construction of realities in organisms and robots. *Kybernetes, 1/2*, 118-128.

Ziemke, T. & Sharkey, N. (2001) A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life. *Semiotica, 134*(1- 4), pp. 701-746.

Ziemke, T. & Thieme, M. (2002) Neuromodulation of Reactive Sensorimotor Mappings as a Short-Term Memory Mechanism in Delayed Response Tasks. *Adaptive Behavior, 10*(3/4), pp. 175-199.



Forsythe, K. (2008). *Meditation 1* (detail). 15 cm x 22 cm, acrylic on canvas.