New Ideas in Psychology xxx (2009) 1–16

Contents lists available at ScienceDirect

New Ideas in Psychology

journal homepage: www.elsevier.com/locate/newideapsych

Towards the naturalization of agency based on an interactivist account of autonomy

Argyris Arnellos*, Thomas Spyrou, John Darzentas

Department of Product & Systems Design Engineering, University of the Aegean, Syros, Greece

PsycINFO classification:
2140
2340
2343
2360
2630
2820

Keywords:
Autonomy
Agency
Naturalization
Functionality
Meaning
Interactive representations

A B S T R A C T

This paper attempts to provide the basis for a broader naturalized account of agency. Naturalization is considered as the need for an ongoing and open-ended process of scientific inquiry driven by the continuous formulation of questions regarding a phenomenon. The naturalization of agency is focused around the interrelation of the fundamental notions of autonomy, functionality, intentionality and meaning. Certain naturalized frameworks of agency are critically considered in an attempt to bring together all the characteristic properties that constitute an autonomous agent, as well as to indicate the shaping of these notions/properties during the development and the evolution of its agential capacity. Autonomy and interaction are proved to be key concepts in this endeavor.

1. What does it mean to naturalize?

There are many different kinds of naturalism, but almost all of its adherents and especially those who advocate epistemological naturalism (viz. Quine, 1969, "Epistemology Naturalized", see e.g. Feldman, 2006), have as a common point the fact that they provide different answers from those of traditional epistemologists, to crucial epistemological questions such as the source of particular beliefs, when particular beliefs are valuable, and how to form such beliefs regarding a certain inquiry. As a result, explanation within naturalized epistemology is understood in a different sense than in traditional epistemology, on which inferences are justified by a priori (and often universal and ontologically valid) beliefs, and by observed data on the behavior of the system.

* Corresponding author.
E-mail addresses: arar@aegean.gr (A. Arnellos), tsp@aegean.gr (T. Spyrou), idarz@aegean.gr (J. Darzentas).

0732-118X/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.newideapsych.2009.09.005

Naturalization requires the justification of an explanation based on natural relations or interactions. Such an explanation is not just an observer's adaptive strategy for interpreting the behavior of other systems, in terms of the observer's beliefs and desires, as Kampis (1999) suggests, it is also an attempt to look inside the system and try to understand and explain how it works. This seems to be a valid strategy for naturalism, as in such cases, the respective explanations can be objectively verified. For this reason, as Etxeberria and Moreno (2001) argue, there is a beneficial influence from the 'natural' sciences to the 'human' sciences, where the benefit is advancement towards a better explanation of the phenomenon under investigation.

One should also keep in mind that the history of science is not just a mere accumulation of empirical data regarding the observed phenomenon that fits to the curves of an analytical model. More importantly, it also involves the radical transformation of the observer's understanding regarding something that he has already observed. Faith (2000) describes these transformations as further naturalizations made possible by the discovery of new and different mechanisms underlying the respective phenomena. Since science is inherently progressive, the resulting explanatory principles and rules regarding those phenomena should not be bounded. On the contrary, they should be mapped onto this progressive nature and hence, naturalization has no end or a specific and discrete final state, but it is an ongoing and open-ended process of scientific inquiry. In other words, naturalization is the continuous formulation of questions regarding a phenomenon considering the quantitative but also the qualitative progress of science regarding notions and beliefs pertaining to this phenomenon, and aiming towards a better understanding and modeling of this phenomenon.

Thus described, naturalization can be considered as the cornerstone of interdisciplinary research, a wider paradigm in which contemporary researchers and scholars should try, in general, to analyze and define in an open-ended manner the notion of agency and in particular, to understand and explain the complex cognitive phenomena in which an agent is involved. In such a naturalistic domain of explanation, agency may acquire many different scientific descriptions and explanations, as long as the theoretical notions used in these descriptions and explanations are naturalistically valid (with reference to contemporary scientific findings); as long they are not based on metaphysical assumptions and a priori judgments. For example, the well-known problem of intentionality and of the intentional behavior of an agent cannot be taken to be answered with a simple appeal to God or to some mysterious dark forces with a metaphysical grounding. At present, this would not be classified as a naturalistic explanation of intentionality. On the other hand, agents act so as to reach out towards the world and as Kampis (1999) argues, the acts and thoughts of an agent always have a target, an object, a referent, or in general, a state of affairs. This provides agency with a certain directedness, and the attempt to explain this directedness by postulating mental states appears to serve the purpose of naturalization well, if not well enough. An explanation based on the existence of neural mechanisms that somehow manage to be in a state that underwrites intentional behavior is a step forward towards a more naturalized explanation. But here, too, the process of naturalization remains incomplete, as this 'magic' state might well be immanent and independent of the environment of the agent, or simply emergent under certain contextual conditions.

As we have claimed, the naturalization of agency requires the explanation of how an agent does something and hence, the naturalization of intentional behavior is not just a matter of *what* an agent does, but also, *how* it does it. Since agents seem to evolve within a dynamically changing environment, while they are always engaging in intentional interaction with it, more valid inquiry within the quest for naturalizing agency would likely be to analyze how the ongoing existence of an agent justifies their intentional behavior. Of course, at this point one may even ask if intentionality is the right property for a naturalized explanation of agency to begin with.

Intentionality is one of the fundamental properties of an agent, but we will show, it is not the only one. In the next section, some familiar definitions of agency will be canvassed in an attempt to locate the proper starting point for a naturalized explanation.

2. Defining agency with a view towards naturalization

Franklin and Graesser (1996) provide an interesting review of many characteristic definitions of the term 'agent'. Some of them (those that seem closest to a stronger, living-system-centered notion of agency) are quoted below.

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

- Russell and Norvig (1995, p. 33) state that: "An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors."
- Maes (1995, p. 108) states that: "Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed."
- Hayes-Roth (1995) states that: "Intelligent agents continuously perform three functions: perception of dynamic conditions in the environment; action to affect conditions in the environment; and reasoning to interpret perceptions, solve problems, draw inferences, and determine actions."
  - Wooldridge and Jennings (1995, p. 2) "Perhaps the most general way in which the term agent is used is to denote a hardware or (more usually) software-based computer system that enjoys the following properties: autonomy: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state.
  - Social ability: agents interact with other agents (and possibly humans) via some kind of agent-communication language.
  - Reactivity: agents perceive their environment, (which may be the physical world, a user via a graphical user interface, a collection of other agents, the INTERNET, or perhaps all of these combined), and respond in a timely fashion to changes that occur in it.
  - Pro-activeness: agents do not simply act in response to their environment, they are able to exhibit goal-directed behavior by taking the initiative."

There are some interesting points in all these definitions. In the first place, all authors conceive and describe an agent in anthropomorphic terms, independently of the agent's area of application (i.e. information system, robotic systems, web-interactive systems, etc.). They all speak of an agent as perceiving its environment, reasoning in order to interpret its perceptions and to draw inferences, acting in the environment, solving problems, communicating with other agents and thus, socializing, etc. It also seems that they all, either explicitly or implicitly, accept a kind of pro-activeness in an agent. And last, but not least, all these definitions ascribe self-rule and independence to the agent.

There is nothing wrong with these features (although Maes seems to be unjustifiably willing to ascribe autonomy to a computational agent), but one should not forget that our present aim is to naturalize agency, so we must determine whether these are the most fundamental, or the most easily naturalized, of an agent's properties. So far, most of the attempts to build artificial agents have been based on the view that concepts such as autonomy and properties such as perception and action are discrete modules of a bigger system and that if one manages to build and integrate these modules, agency will result.

But concepts such as autonomy and pro-activity, and even 'simple' properties such as perception and inference, are not all-or-nothing propositions. A cat, a parrot and a man can all be fully described by the definition of Wooldridge and Jennings, although the degree of autonomy, pro-activity, and socialization through perception and action is not the same in these three agents. With some imagination, even a rock might be made to fit this description. Agency does not come in an all-or-nothing package and there are various levels of agency in the biological realm. Agents are not static things, but complex systems interacting with dynamic and complex environments, and, therefore, exhibiting a dynamic nature in themselves. There are some dynamic and incremental conceptual and material ingredients that are complexly integrated to constitute an agent to varying degrees and at various evolutionary. Therefore, different agents exhibit different degrees of agency at a specific point in time, and the same agent may also exhibit different degrees of agency at two different points in time.

Keeping naturalization in mind, one may conclude that a complete definition of the term 'agent' is out of the question, and any efforts in this direction should express agency as a capacity with an incremental and an evolutionary character. In order to pursue such definition, we will try to modify Kampis's (1999) evolutionary definition of agency, which comes as a list of somewhat *ad hoc* properties of an agent, in such a way that the suggested definition is more susceptible to a naturalized analysis. Therefore, since the ability to act upon an environment in order to achieve goal-oriented effects is proper to a cognitive agent, we suggest that a strong notion of agency calls for:

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

interactivity, that is, 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 engage in a goal-oriented interaction driven by beliefs and desires; and
autonomy, the ability of an agent to function/operate intentionally and interactively based on its own resources.

This definition does not try to state what an agent is or what an agent should do, instead mentioning three nested capacities that an agent should exhibit over the course of its evolutionary development. According to this definition agency requires interactivity, which in turn implies action upon the environment. This action is not accidental but intentional, as it is a goal-directed and 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. These three properties seem to be quite interdependent, and there is no possibility that any one of them may be increased qualitatively in isolation from the others.

However, notions such as autonomy, intentionality, belief, goal-orientation, cognition, etc. are philosophically loaded and controversial. This theoretical piggy-backing may prove a terminological and conceptual burden, especially considering that such an analysis is meant to serve to inspire and to guide the creation of artificial agency. With a critical perspective in the domain of complex systems research, Collier (1999) suggests that there is a very interesting interdependence between these three 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 system's *autonomy* during its purposeful *interaction* with the environment.

This conceptual interdependence places some interesting constraints on the capacities that contribute to agency and sets some requirements in terms of the properties that an agent should exhibit independently of its degree of agency or cognitive capacity. These properties and their interdependence are characteristic 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 – the 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 cognitive system as a whole (see e.g. Ruiz-Mirazo & Moreno, 2004). Meaning, if it is not to be considered as an ascription by an observer, should be linked with the functional structures of the agent. It should guide the constructive and interactive processes of the functional components of the cognitive system in such a way that these processes maintain and enhance the system's autonomy. In this perspective, the enhancement of autonomy imposes certain goals on the agent itself and hence, the intentionality of the agent is guiding its behavior through meaning.

It should be noted 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 the defining properties of an autonomous agent that may act 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 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 create new meanings that will further enhance its autonomy. Moreover, agency has an interactive and a goal-oriented character, which results from the interactivity and the intentionality of the respective cognitive system.

In the rest of this paper, an attempt is made to ground these properties in naturalistic frameworks of cognitive agency, and to describe their emergence during the development and the evolution of a cognitive agent. Autonomy will be the key concept in this endeavor.

3. Second-order cybernetics: agency as self-organization

In general, within the framework of second-order cybernetics (von Foerster, 1981, 2003), agency is considered as the continuous process of modification of the intentional behavior of the system

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

through the constant alteration of its organization. This conception is radically different from the traditional cognitivistic frameworks (see e.g. Newell, 1980), on which agency is the gathering and assembly of well-defined representations about the environment. Consequently, the focus on agency is shifted from a serial and static mapping between the internal and the external of a cognitive system to the coupling of constant and parallel structural processes of the system with its environment.

It should be noted that the broader framework of second-order cybernetics and self-organization opens the black box of the cognitive system, provided that it offers a 'mechanism' that supports the behavior of a cognitive agent. At the same time, it neither presupposes nor predetermines the 'vocabulary' with which each cognitive system chooses to shape the various states of affairs in the environment and to act about and towards them. In other words, each cognitive system constructs its own meaning based on its actions in the environment. As such, the approach of second-order cybernetics and of self-organization in general is an interesting candidate for a non-circular way of describing and explaining the reference of the thoughts of a cognitive agent. The non-explicit or in a non-instructive participation of the environment in the shaping of the cognitive process and of the meanings generated in the self-organized system endows the system with a high degree of autonomy. As has already been mentioned, autonomy has the primary role in the establishment of a naturalistic framework for the analysis, explanation and modeling of the emergence and further development of meaning in a cognitive system—the emergence and development of agency.

In the following section, we provide a descriptive analysis of the framework of second-order cybernetics, indicating the main characteristics of the self-organizing approach to agency.

3.1. Organizational closure and self-reference for self-organization

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 creating new meanings in order to interact (Luhmann, 1995). The self-referential loop can only exist in relation to an environment. In contrast with classical system-environment models, the external control of a cognitive system's adaptation to its environment is replaced by *systemic* and *operational/organizational closure* (von Foerster, 1960, 1981).

Due to that closure, the self-reference of an observation creates *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, which von Foerster (1981) calls *eigenvalues*, postulating 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 by forming new and more complex organizations.

Given such organizational closure, environmental complexity is based solely on system observations, thus, system reality is observation-based. As von Foerster (1976) has argued, the results of an observation do not refer directly to objects in 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, p. 266). Thus, each new function based on observations is a *construction*, it is an increase in the organizational and cognitive complexity of the agent. This process of emergent incrementation of order through the internal construction of functional organizations and simultaneous classification of the environment is a process of *self-organization* (von Foerster, 1960, 1981).

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

3.2. Agency and autonomy in autopoiesis

Maturana and Varela have made a very interesting attempt to capture the essence of the processual and constructivist nature of agency through the introduction of the notion of autopoiesis (Maturana & Varela, 1980; Varela, Maturana, & Uribe, 1974). At the same time, Maturana and Varela (1973) introduced the notion of autonomy as a central concept in the study of biological, cognitive and adaptive systems.

Within the theory of autopoiesis, Varela (1979, 1992), and Varela and Bourgine (1992) defines the concept of autonomy as an abstract kind of organization, as self-maintained, self-enhanced and self-regulated systems dynamics originating from a *network of processes with a high degree of recursivity* that produces and maintains internal invariances in the face of internal and external perturbations. As such, autonomy is a self-defining process that establishes the uniqueness of a system as differentiated from all other surrounding processes. This is a rather abstract conception of autonomy, but nevertheless, Varela argues that such autonomy is realized in different biological scales and domains. For instance, in the framework of autopoiesis life is defined as a special kind of autonomy, the one of achieving autopoiesis in the physical space. Cognitively driven behavior is the result of a higher level of autonomy, wherein 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. This is the reason why in this framework the naturalized understanding of the cognitive process is indissolubly connected with the phenomenon of life and of being alive in general. In the autopoietic framework, the capacity for agency requires both the capacity of being alive and that of acting in the world, each of which is established by the respective kind of autonomy.

For Varela, the second-order cybernetic notions of closure and self-reference, as well as the organization-centered perspective of the systemic view play a most important role in the constitution of the notion of autonomy. Specifically, 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)

It appears that for Varela, autonomy is equivalent to the notion of self-referentiality, which in turn 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 to the self-maintenance of the autopoietic system, and especially to the production of its active components and the effective alteration of its boundary conditions for the maintenance of homeostasis.

Varela's conception of autonomy has often been criticized, mainly due to its emphasis on the absoluteness of organizational closure, and the secondary role it ascribes to both environment and the interactive aspects of the system (see e.g. Collier, 2000, 2002). However, the conception of autonomy derived from autopoiesis stresses two really important aspects of the notion of agency, and prepares the way for more elaborate system-theoretical approaches. Specifically, agency needs cohesion and complex boundary conditions.

3.3. Cohesion and constructions via process closure for self-organizing agency

According to the theory of autopoiesis, what defines life is a global network of relations that establishes self-maintaining dynamics, where living systems are constituted by the capacity for action in the service of self-maintenance. Practically, this means that the activity of the system consists in the constant regeneration of all the processes and components that constitute the system as a functional unit. Any circularity here is merely superficial, given the self-referential, organizationally closed nature of such an autonomous system. The internal productive relations acquire a cohesive functional meaning in a collective way, since they contribute to the overall maintenance of the system. Actually, in the kind of self-organization implied in autopoiesis the whole and the parts are correlated to each

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

other in a highly reciprocal manner. Indeed, autopoiesis, strongly influenced by the systemic roots of second-order cybernetics, offers a mechanism for self-organization that disregards the classical mechanistic opposition between the constituent parts and the global properties of the autonomous system (Luisi, 2003).

However, this systemic pattern of organizational (functional) dynamics is observed in every self-organizing system. Collier (1988) and Collier and Muller (1998) had called this pattern of organizational dynamics *cohesion*, an inclusive capacity of autonomous systems, indicating the existence of a logical closure of the relations among the elements of a system that contribute to its maintenance. Cohesion is not an epiphenomenal property, but the result of causal interactions among the components of the system in which certain capacities emerge. As such, the respective components are constituents of the system itself.

Cohesion is a property embedded in the dynamic organization of a self-organizing system, one whose presence leads to further organizational complexity. As such, it can only be explained by reference to the causal roles that constituent components and the relations among them acquire in the dynamic organization of the system. This kind of organizational complexity requires systems that are thermodynamically open and function in far-from-equilibrium conditions (Collier & Hooker, 1999; Ruiz-Mirazo & Moreno, 2004; Ruiz-Mirazo, Pereto, & Moreno, 2004). Given only these thermodynamic conditions, we may assume that the management of the energetic aspects of such a system characterizes the degree of agency. Such is the case for systems that present minimal or/and marginal self-organization, where the group of components are at the critical 'edge of chaos' (see e.g. Kauffman, 1993).

However, in self-organizing systems with a certain degree of stability, such as autopoietic systems, the degree of autonomy is more appropriately characterized in relation to the organizational rather than the energetic aspects of system processes. Such systems exhibit long-range correlations between different processes, and as Collier (2007) has stressed, since there is an internal need for the coordination of processes in order for them to achieve viability, one should expect to find in an autonomous system a holistic organization in which organizationally/operationally open aspects of lower levels are closed at higher organizational levels. This *constructive* type of autonomy (Ruiz-Mirazo & Moreno, 2000) requires what Collier (1999) describes as *process closure* (in accordance with organizational/operational closure). In such autonomous systems, there are internal constraints controlling the internal flow of matter and energy, allowing the whole system to acquire the capacity to carry out the processes that contribute to its viability.

In this level of autonomy, agency is considered as a process of constant alteration of the intentional behavior of the self-organizing system through the continuous modification of its functional organization. In other words, a self-organizing system is able to both establish and change its functionality in order to interact with an environment. This provides the self-organizing cognitive system with a kind of autonomy that is not supported in the classical symbolic/cognitivist frameworks, since in the latter, all functional change must be externally imposed.

Furthermore, the nature of systemic organizational- and process closure means that all the interactive alternatives of the cognitive system are internally generated and their selection is an entirely internal process. Therefore, such autonomous cognitive systems must construct their reality by using internally available structures. One should note that the respective self-organized structures (eigenvalues) are specific to functional particularities of the cognitive system. Specifically, the functionality of the cognitive system 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.

The critical question for this view of agency has to do with the source of the intentionality of the cognitive system. In general, it locates intentionality and especially, the endogenous production of *purpose* at the level of the origin of life and of biological functionality. This should no longer come as a surprise, since in a framework of agency based on self-organization a cognitive system is first and foremost a living system. 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 a strong notion of embodiment based on the dynamics of the

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

functional organization of the cognitive system and it is stands in stark contrast to the almost disembodied nature of a purely symbolic system.

This kind of agency requires cohesion for internal constructions via process closure. The respective model shows a satisfactory level of naturalization, as it has managed to ground agency in the functionality of the self-organizing system by identifying the former with autonomy and introducing some requirements for the latter. However, this model is not adequate—for the purposes of this paper, it remains insufficiently naturalized. Cohesion via process closure is a necessary but not a sufficient condition for agency, as rocks and crystals show great degrees of cohesion, but do not exhibit any significant intentionality and as such, cannot be considered agents. Agency comes in many degrees and different levels in nature, but almost everybody would agree that living systems are quite different from rocks. Indeed, they are quite different in terms of their degree of autonomy, especially as regards to their degree of disentanglement from the environment. An agent with only these characteristics cannot go too far in advancing its evolutionary and developmental horizons, as it is too tightly coupled with the environment. At this point, one can speak of systems being at the threshold of autonomy, and thus exhibiting merely a reactive type of agency.

Reactive agency grounded in cohesion, based only on the process closure that drives internal construction, is not enough for the enhancement (or sometimes even for the maintenance) of autonomy and hence, for the development of broader and more versatile types of agency (in an organizationally and functionally closed and reciprocal system). The reason for this limitation comes from the fact that this specific kind of embodiment and the consequent autonomy do not come gratuitously. Specifically, the self-organizing system can only grasp those aspects of its environment that can be constructed by its limited functional dynamics. Therefore, the meaning constructed by such an autonomous cognitive system is not open-ended. As a result, its functionality cannot support the ongoing alteration of its dynamic organization in order to respond to new and unforeseen processes operating across its boundaries between system and environment.

On the other hand, active agency is open-ended and emerges out of intentional and mostly ill-defined goals and purposes of cognitive systems. This means that the anticipatory content of each self-organizing system interacting with the environment should be open to revision and evolution, in order to reflect both this ill-definedness and open-endedness. In this way, the self-organizing system will have the ability to engender new functions that will be directed towards new goals and hence, the new functionality will contribute to the autonomy of the system in new ways (Collier, 1999). Therefore, active agency cannot be solely a matter of internal constructive processes and process closure. The need for open-endedness calls for a rich and versatile interaction of the self-organizing system with the environment, while, the functional aspects of such a constructivist embodiment and its anticipatory content call for advanced and efficient mechanisms for controlling and managing these interactions.

Before considering the advantages of the dynamic opening of an autonomous system to the environment through the development of complex interactive capacities, we must describe the complex boundary conditions that allow for the emergence of agents capable of developing such capacities.

4. From agency as self-organization to agency via enhanced interactive capacity

4.1. Complex boundary conditions and the formation of inside/outside asymmetries

In Section 3.1, we note that any system (defined and considered under the more general framework of second-order cybernetics) makes a distinction between the components that constitute itself, and the elements that form its environment. In making this distinction, the system observes both domains (internal and external) by observing of its own boundaries.

The resulting qualitative and quantitative imbalance indicates an asymmetry between the system and its environment. 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 organizationally differentiate the system from its environment, and furthermore they specify its autonomy and its identity.

Hoffmeyer (1996, 1998) strongly argues that the secret of life and the development of agency are hidden in this asymmetry. Hoffmeyer (1998) sets out from the somehow mysterious answer (especially

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

when articulated solely in terms of second-order cybernetics) that von Foerster provides to Vijver's question (see Vijver, 1997) concerning to the fundamental problem of second-order cybernetics, that it attempts to develop a theoretical conception of the observer without any conception of the subject (as an individual agent). Von Foerster replies that in this case one needs:

> an epistemological salto mortale, because the moment you open your mouth *you* open the mouth, but to identify therefore what is coming out of the mouth which has been opened is reflecting the open mouth (Vijver, 1997, p. 5, emphasis in the original).

In general, von Foerster's reply points to the need for self-reference and closure in any theoretical conception of the subject, i.e., of an autonomous system, but he does not say anything else regarding the implications of these properties or the ways they might be achieved. Revisiting this issue, Hoffmeyer (1998) stresses the importance of 'inside exterior' and 'outside interior' as two conceptual categories that are reflected in the real world at the relation between inside (the system itself) and outside (the environment) of a system. He argues that the quest for the origin of life is identified with the quest for the origin of the environment. In this way, living systems are interwoven with the environment, although they are asymmetrically differentiated. From the system's point of view, the environment is anything outside of it, while from the environment's point of view, systems are considered as encapsulated entities within the environment. He also suggests that this dichotomy between the internal and the external aspect the asymmetry should in no way be considered as absolute. Hoffmeyer justifies this claim by noting that when the dichotomy is closely examined, we must conclude that the external part of a given asymmetry tends to be an internal part of something else. In addition, something belongs to the external part for a given span of time, may belong to the internal part at some other time.

According to Hoffmeyer, this seems also to be the key to the problem of the asymmetry between system and environment. He suggests that life and agency are constructed upon a fundamental asymmetry, but not on an absolute one. Biologically speaking, Hoffmeyer suggests that such an asymmetry is produced via a semi-permeable membrane. At this level of analysis, and considering the requirement for a naturalized explanation, this should not come as a surprise. As a matter of fact, (Ruiz-Mirazo & Moreno, 2000, 2004) have argued in detail that the boundaries of self-organizing systems have a functional basis of a chemical nature, as they are the result of a productive organization and activity of the self-regulating and self-modifying processes of their systems. This self-regulation aims at the maintenance of the system. This active relation between the boundary of a self-organizing system and the recursive production processes of its constitutive components is also exemplified in the autopoietic model, but with an emphasis on the absoluteness of the control and constrain of the flows of energy and matter into the system from the environment. As also suggested by Hoffmeyer, this relation is a relation of regulation, hence, it cannot be absolute.

However, for present purposes the material basis of the complex boundary supporting the asymmetry need not concern us. What is really important are the logical implications of such a boundary and the implied asymmetry. As Hoffmeyer (2001) has pointed out, the sterile and mechanistic view of biochemistry must be abandoned in service of the attempt to understand the relations between inside and outside as *measurements* with a *referential* nature. What Hoffmeyer means is that one should try to understand the regulatory relations between an autonomous system and its environment based on the *interpretive interactions* that take place across the boundary of the system. The resulting interpretations result in further actions in the system, hence in any potential contribution to its maintenance. Consequently, these semi-permeable membranes are not just physical borders, but *interfaces* that function as active regulators of the interpretive interaction with the environment and as such, they have a very important role in the maintenance and the further development of the organization of the cognitive system, and of course, of its agentive capacity.

4.2. Interaction closure and the emergence of functional norms

The interpretive asymmetry described in the previous section implies the view that the enhancement and evolution of autonomy and agency require ever more complex interpretive interactions with the environment. Consequently, the interactive opening of the system to the environment must be

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

considered the most important point in its evolution towards genuine agency, as it marks the point beyond which the system can create different meanings inside itself. This meaning is grounded in the functionality of the system, with immediate implications for its self-maintenance and further development.

Specifically, as an agent interacts, its domain of interpretation differentiates between two kinds of interaction, namely between *functional* and *dysfunctional* (Moreno & Barandiaran, 2004). The former corresponds to interactions that are integrated into the functional organization of the agent and contribute to its self-maintenance. The latter corresponds to the interactions that cannot be properly integrated in the functional organization and hence, they either fail to contribute to or actually disturb the self-maintenance of the system.

Therefore, such autonomous systems do not only exhibit process closure, but also *interaction closure* (Collier, 1999, 2000), a situation in which the internal outcomes of the interactions of the cognitive system with its environment contribute to the maintenance of the functional (constructive/interactive) processes of the system that are responsible for these specific interactions. As it we noted in Section 3.3, cohesion via process closure is not a sufficient measure for autonomy and hence for genuine agency. But cohesion via process and interaction closure is exactly what distinguishes truly autonomous systems from other kinds 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 in the environment. This capacity for meaningful critique regarding the 'good' and the 'bad' with respect to the maintenance of the system is a normative one. Self-maintenant systems that exhibit normative functionality are truly autonomous systems that present genuine agency.

Functional norms, are emergent in system's interactions with the environment. Specifically, they are internal constructions that attribute binary values to the processes or/and interactions of an autonomous system. The binary character does not imply explicitness; to the contrary, the higher the degree of autonomy and agency of the cognitive system, the higher the degree of abstraction of the concepts to which some of its norms can be applied. This means that even though norms are constructed by the autonomous system itself, there are still many cases in which the success or failure of their interactive satisfaction is not immediately recognizable by the system. When this happens the system is not sure about the possible outcome of an interaction, and it chooses to resort to its anticipations. As we discuss in the next section, the naturalistic requirement for an explanation of the constructive and interactive aspects of normative functionality, i.e., of the efficient control and management of the constructive/interactive capabilities of an autonomous agent, calls for the introduction of interactive emergent representations.

4.3. The need for interactive representations in autonomous agents

The case of representational content and its role in the functionality of autonomous agents is highly controversial (see e.g. Bickhard, 1993; Bickhard & Terveen, 1995 for a detailed analysis of the problem).

The internal constructions with which the self-organizing system classifies the environment and acts on it are not themselves representations of the environment. As von Glasersfeld (1995) argues, these constructions are instead *re-presentations* generated by the cognitive system in its embodied interaction with the environment. In second-order cybernetics, memory is understood as a process of re-presenting and re-membering by bringing past experiences into the present (von Foerster, 1969). Hence, *re-presentations* refers to the self-organized dynamics by virtue of which a previous construction is re-constructed (re-presented) from memory, given that there is some sensory interaction (perturbation) with the environment.

In general, in the context of second-order cybernetics, the notion of representation as encoded information in exact correspondence with the aspects of the environment that are supposed to be represented, is totally rejected. Second-order cybernetic system models admit no functional usefulness to representations and they regard information only as socially ascribed to a process by an observer (Maturana & Varela, 1980; von Glasersfeld, 1995).

This rejection somewhat constrains the autonomy of a self-organized system to its internal dynamics. In addition, there remain some cases for which representations are really required. Hence,

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

Clark and Toribio (1994) argue in favor of 'representationally hungry' phenomena, which occur mostly in the daily action of cognitive agents. In a more inclusive manner, Bickhard and Terveen (1995) note some characteristic cases in which a kind of interactive representation that makes possible the internal detection of error is necessary for the successful functioning of the cognitive system. In these cases, "the processing in the system must be potentially controllable, at least in part, by system error…" (Bickhard & Terveen, 1995, p. 210). Such cases appear in goal-directed interactions, "when system implicit anticipation of the courses and outcomes of interactions cannot be assured" (Bickhard & Terveen, 1995, p. 211) and in learning processes, as "Learning cannot be fully successfully anticipatory—if it were, there would be nothing to be learned. Learning must involve the possibility of error, and such error must be functionally detectable by the system itself so that the learning can be guided by it" (Bickhard & Terveen, 1995). Another case for which interactive representations are needed is when there is more than one possible course of interaction for a specific environment and the system must choose among them on the basis of each anticipated outcomes of these interactions (Bickhard, 2001; Bickhard & Terveen, 1995).

It is apparent that the higher the degree of agency, the deeper the implications of autonomous cognitive system-environment interaction, since some norms of higher-level agents can be satisfied with more than one interactive strategy. Additionally, different interactive strategies may have more than one independent consequence. Selections cannot be made through simple triggering, but require some more complex process in determining the course of the interaction. Of course there are some cases, where particular sensory interactions are known to provoke specific responses, especially in well-defined anticipations, where there is no need for the cognitive system to be aware of the subsequent internal outcomes. As was discussed in Section 3.3, these cases are characteristic of purely reactive systems and their respective models, and cannot provide a naturalistic explanation for intentional and purposeful interaction of the autonomous cognitive system. Something is needed that will justify the relation of internally self-organizing structures of the autonomous cognitive system to particular aspects of their interactions with particular states of affairs in the environment.

4.4. The emergence of interactive representations in autonomous agents

Such 'informational' internal states, which refer to certain conditions of the environment, must have an embodied and situated character (Moreno, Umerez, & Ibanez, 1997) in order to be able to ground the representation to the context of the situated interaction between autonomous agent and the environment. Indeed, considering the functional closure of a self-organizing system, its constructions can be seen as internal *in-formational* patterns, which have nothing to do with the transference of ontological information from the environment to the system. As long as this internal construction permits an agent to survive, at least in this specific environment, and hence, to maintain or even enhance its autonomy, this construction should be considered as a *representation* of the situated interaction of the agent with the respective environment.

Bickhard (1993, 2000, 2001) exemplifies this situation by postulating a recursive self-maintenant system, which is a self-organizing system that has more than one means at its disposal to maintain its capacity for self-maintenance under various environmental conditions. This is a self-organizing system that avoids going to equilibrium by continuously interacting with the environment, in the process determining the *appropriate conditions* for the success of its functional processes. 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 such that its functionality is immediately related to its autonomy, by virtue of 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.

In this way, the overall functional closure (process and interaction closure) of the cognitive system is guided by its autonomy, in the sense of the former's contributing to the maintenance of the latter, while its intentionality derives from this specific normative functionality, which in turn is directed towards the primary purpose of maintaining self-maintenance. What is still missing is meaning, on the basis of which the cognitive system decides which of the available functional processes it should invoke, in

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

order to successfully interact with a specific environment, that is, in order to fulfill its goal. But, where exactly is this meaning to be found?

Bickhard argues that such an autonomous system should have a way of differentiating the possible environments with which it interacts, and a switching mechanism in order to choose among the appropriate internal functional processes that it will use in a given interaction. The differentiations are implicitly and interactively defined, as the internal outcomes of the interaction. These differentiations can occur in any interaction, and the outcome of the interaction depends on the organization of the participating subsystems and of the environment. Bickhard emphasizes that such differentiations create epistemic contact with the environment, but they do not carry any representational content, thus they are not representations by themselves. Rather, they are indications of the interactive potentiality of the functional processes of the autonomous cognitive system itself.

More specifically, the role of these differentiations is twofold: *a*. they indicate the range of interactions that are functionally available to the cognitive system to use in this specific environment, i.e., they indicate which further interactions might be possible or appropriate (Bickhard, 2000), by virtue of at least contributing to the maintenance of the autonomy of the cognitive system; *b*. they implicitly presuppose the environmental properties that would support the success of the functionally indicated interactive processes. In other words, such differentiations functionally indicate that some type of interaction is available in the specific environment and hence, implicitly presuppose that the environment exhibits the *appropriate conditions* for the success of the indicated interaction.

In this model, such differentiated indications constitute *emergent representations*. The conditions of the environment that are functionally and implicitly presupposed 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 with respect to the differentiated environment. This content emerges in the interaction of the system with the environment and it corresponds to the *implicitly defined supports* of the functionally indicated interactive process (Bickhard, 2000).

This content may be in error, which means that the respective dynamic presuppositions may not hold (i.e. the environment may not satisfy the presupposed conditions). But this error will be functionally detectable by the cognitive system itself, since it will be functionally evaluated on the basis of contribution to the maintenance of the autonomy of the system – the indications of the content are embedded in the functionality of the system. Hence meaning is produced by the functional evaluation of representational content, internally in the autonomous cognitive system, but in the interaction of the system with its environment. It is in this way that meaning is a prerequisite of, and contributes to the maintenance of the autonomy of the cognitive agent during its intentional and purposeful interactions.

From this perspective, each referential state of the autonomous cognitive system should be considered as situated in the context of self-organized *in-formational* structures, as these are internally constructed due to its functional/organizational closure. In particular, these in-formational structures determine the intentional and purposeful interaction of the autonomous cognitive system based on the range of indicated organizational forms they can support. Therefore, these in-formational structures indicate the representations that emerge, and can only be defined in the context of the interaction of the autonomous cognitive system with the environment. In other words, any representational functional organization is an emergent product of the interaction between the autonomous system and its environment. Hence, in an autonomous system, functionality provides intentionality simply because its functional structure carries, during the interaction, potentially reliable content about the environment.

So far, we have tried to provide a naturalized explanation of agency through the analysis of the way an autonomous agent emerges and develops as it interacts with the environment. At this point, an autonomous agent uses its own functions in order to intentionally interact with a dynamic environment. In other words, agency is conceived as identical to autonomy, since it is viewed as a qualitative measure for the interactive potentiality a self-maintaining system is capable of managing. This leads to ever-greater degrees of disentanglement from the environment, and hence to agents that exhibit ever-greater degrees of autonomy.

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

In the next section, we discuss ways in which this context of purposeful interaction can be further enhanced in the face of complicated goals, further contributing to the agent's autonomy and in general, to its agential capacity. As one might expect from the analysis in Section 4.2, anticipations will play the central role.

5. Agency based on dynamic anticipations: enhancing the capacity for interaction

5.1. Anticipation and functionality

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 sustain 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; purely reactive systems are thus incapable of anticipative functionality), hence, anticipation is immediately related to the meanings 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 interactions 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, of 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, 2000). Anticipation in turn, is goal-directed. As a matter of fact, anticipation almost always requires functionality, which is, by default, a goal-oriented process. From 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 autonomous self-organizing systems 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 whose potential error is detectable by the system itself, since such anticipation is embedded in the functional context of a goal-directed system. This type of anticipation is very different from that supported by the cognitivist models of representation, which attempt to map the environment to their past decisions. Here, the activity is future-oriented and it can be mistaken, if the chosen interactive strategy does not internally yield the desired results, or if the respective environment does not support the type of interaction that would lead to the anticipated internal outcome. This is a naturalized account of interactive anticipation.

5.2. Dynamic anticipation enhances autonomy and agency

As argued above, agency is open-ended and emerges out of the intentional goals and purposes of self-maintainant autonomous systems. The anticipatory content of such autonomous systems should be open to revision and evolution. With respect to the dynamic and future-oriented type of anticipation described above, each autonomous agent should have the capacity for anticipative interaction with the environment, in order to achieve the closure conditions that will further contribute to its autonomy.

As we have said, the only way for an autonomous system to enhance its autonomy is by pursing its goals through the construction of even more adaptive representations. In general, the more the representational content of an autonomous system is evolved the more dynamic its anticipative structures become (Bickhard, 2001; Collier, 1999). This has a positive effect on the anticipatory capacity of the autonomous system and in its capacity to evaluate its future interactions. The increase of the system's capacity for dynamic anticipation expands what Christensen and Hooker (2000) call the *anticipatory time window*, which provides a certain degree of directionality (Christensen & Hooker, 2002) in the goal-directed interaction of the autonomous system. Overall, these capacities result in the emergence of new cognitive abilities for the autonomous system, thus implicitly increasing its interactive autonomy.

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

Nevertheless, no matter how large the window of anticipatory interaction may be, all possibilities and selections regarding the outcomes and the ill-defined consequences of the design process cannot be inherent in the innate organization of each autonomous agent. One possible solution is for the agent to evolve learning capabilities. This would provide a way to expand its capacity for dynamic anticipation and its ability to evaluate possible interactions. The agent becomes less dependent and more sensitive in regarding its contextual interactive potential. It improves its ability to recognize its environment, evaluate conditions, and formulate goals regarding the problem. It thus develops an infrastructure better suited to anticipating the possibility of success in emergent interactions with the environment. Structural coupling is strengthened, and the conditions for the emergence of new and more adaptive representational content fostered. Consequently, autonomy is increased.

However, it should be clear that not every external perturbation is useful for an autonomous agent. Only those contributing to the system's closure and, therefore, to the preservation of its autonomy would be selected for further exploitation. Since, in the proposed framework, closure is achieved at the level of differentiations and of the respective emergent representational content, we conclude that autonomy and hence, agency cannot be statically identified. Instead, as Collier (2000, 2002) suggests, they have an incremental nature.

Hence, agency should be considered an anticipative, future-directed property. It is a vital asset directly related to the versatility with which an autonomous system interacting with its environment will internally create adaptive emergent representations directed towards its goals and purposes. Moreover, due to the capacity for directed interaction, an autonomous system, in its attempt to create richer representational structures for its intentional purposes, is continuously interacting with even more complex and dynamic environments and hence, it learns to anticipate, or as is suggested by Bickhard (2001) it anticipates the necessity to acquire new anticipations. Furthermore, the progressively increasing capability of an agent's interactive anticipations creates an intentional capacity. This capacity is not the same as the traditional notion of intentionality considered as the sum of all of a system's representations. Intentionality derives from the agent's functional capacity for anticipative and purposeful interaction, and aims at the enhancement of its autonomy.

6. Conclusion

Naturalization is quite controversial. One has to remain consistent with the natural sciences, while trying not to reduce the whole endeavor to physicalism. For this balance to be achieved, one has to proceed by way of the continuous formulation of questions regarding the phenomenon being naturalized, considering the quantitative but also the qualitative progress of science relating to this phenomenon.

Agency appears to be one of the most complicated capacities that nature presents and the quest for its naturalized explanation is not something trivial. In this paper, we have argued that the naturalization of agency should be concentrated around the naturalized explication of the interrelations between the fundamental properties of autonomy, function, intentionality and meaning, and the way these properties are further developed and enhanced in the evolutionary and developmental history of an agent.

The resulting analysis departs from the systemic framework of second-order cybernetics, on which an agent is considered as a self-organizing system that exhibits self-reference and organization closure. The essence of this kind of agency is exemplified in the theory of autopoiesis, which is the first coherent theoretical conception of agency as congruent with life and of agents as corresponding to living systems in an evolutionary perspective. Autopoiesis and the conception of autonomy to which it gave rise were a very important attempt in the quest to naturalize agency, preparing the way for more elaborate and naturalistically valid systemic approaches. The main contributions of the autopoietic view to the explanation of agency are the need for cohesion among the constituent components of the self-organizing system, and the emphasis on the importance of its complex and active boundary conditions.

Cohesion is a vital property of autonomy and agency when it occurs in the context of both process- and interaction closure. Self-organization entails process closure, which cannot by itself account for genuine agency, but only for reactive systems. However, when combined with complex boundary

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

conditions, self-organization becomes self-maintenance as now the system uses its environment in order to maintain its autonomy. This is achieved as the system begins to interact with the environment on the basis of interpretive evaluations of the respective interactions. This capacity is furnished by the dynamic nature of its boundary, which is formed and maintained by its constructive and interactive processes and which in turn sustains these processes. This functional circularity is the result of interaction closure, which combined with process closure provides a tool for the theoretical distinction between truly autonomous systems and other kind of cohesive systems, such as rocks and stones.

Since not all possible interactions are beneficial to such an autonomous system, its overall functionality acquires a normative character grounded in values emerging within the system and guiding its interactions. As an agent evolves, some of its norms cannot be immediately identified, or satisfied in its current functional organization, and so the system requires some mediation of uncertain interactive potentialities. This mediation is provided by the formation of relevant anticipations with their appropriate representational content. Specifically, in a dynamic environment, an autonomous cognitive system with the ability to maintain the autonomy of its self-maintenance requires the internal generation of representational content that will drive its goal-oriented interactions. The representational content emerges in the respective interactions, and depends upon the dynamic conditions of the environment and of the cognitive system itself. It consists in anticipations that indicate the possibility of future interactions for the cognitive system, and which result in the emergence of new functionality, which in turn is directed towards new goals. The autonomous cognitive agent will continue to interact with the environment in pursuit of these new goals, having as a primary aim the maintenance of its own autonomy.

Finally, the capacity for directed interaction gives rise to the capacity for learning, which prepares an autonomous agent to engage in still more demanding and more complicated interactions with the environment. The prerequisite for learning is that the anticipatory content of the system be open to revision and susceptible to error, where this error is internally detectable by the system itself. These properties are provided by a representational content that emerges in an autonomous system that has the ability to interact with the environment in order to maintain its autonomy. Autonomy drives interaction and profits from it, and as a result enhances the capacity for agency.

References

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*, 111–131.

Bickhard, M. H. (2001). Function, anticipation, representation. In D. M. Dubois (Ed.), *Computing anticipatory systems, CASYS 2000 – Fourth international conference* (pp. 459–469). Melville, NY: American Institute of Physics.

Bickhard, M. H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science – Impasse and solution*. Amsterdam: Elsevier Scientific.

Christensen, W. D., & Hooker, C. A. (2000). Autonomy and the emergence of intelligence: organised interactive construction. *Communication and Cognition – Artificial Intelligence, 17*, 133–157.

Christensen, W. D. & Hooker, C. A. (2002). Self-directed agents. In J. MacIntosh (Ed.), *Naturalism Evolution & Intentionality (Canadian Journal of Philosophy, Supplementary Volume XXVII)* (pp. 19–52). Calgary: University of Calgary Press.

Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese, 101*, 401–431.

Collier, J. (1988). Supervenience and reduction in biological hierarchies. In M. Matthen, & B. Linsky (Eds.), *Philosophy and biology*. Canadian Journal of Philosophy Supplementary, Vol. 14 (pp. 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* (pp. 75–81). New York, Woodbury: American Institute of Physics.

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

Collier, J. (2002). What is autonomy? In *Partial proceedings of CASYS' 01: Fifth international conference on computing anticipatory systems, International journal of computing anticipatory systems*. Vol. 12 (pp. 212–221).

Collier, J. (2007). Simulating autonomous anticipation: the importance of Dubois' conjecture. *BioSystems, 91*, 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. ECHO III Conference, Acta Polytechnica Scandinavica, MA91. In G. Farre, & T. Oksala (Eds.), *Emergence, complexity, hierarchy and organization*. Finish Academy of Technology.

Etxeberria, A., & Moreno, A. (2001). From complexity to simplicity: nature and symbols. *Biosystems, 60*, 149–157.

Faith, J. (2000). Emergent representations: dialectical materialism and the philosophy of mind. DPhil Thesis, University of Sussex.

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005

Feldman, R. (2006). Naturalized epistemology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2006 Ed.). Available from http://plato.stanford.edu/archives/fall2006/entries/epistemology-naturalized/.
Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program? A taxonomy for autonomous agents. *Proceedings of the third international workshop on agent theories, architectures, and languages*. Heidelberg: Springer-Verlag.
Hayes-Roth, B. (1995). An architecture for adaptive intelligent systems. *Artificial Intelligence: Special Issue on Agents and Interactivity, 72*, 329–365.
Hoffmeyer, J. (1996). *Signs of meaning in the universe*. Bloomington, IN: Indiana University Press.
Hoffmeyer, J. (1998). Surfaces inside surfaces. On the origin of agency and life. *Cybernetics and Human Knowing, 5*, 33–42.
Hoffmeyer, J. (2001). Life and reference. In L. M. Rocha (Ed.), *The physics and evolution of symbols and codes: Reflections on the work of Howard Pattee Special Issue of BioSystems, Vol. 60* (pp. 123–130).
Kampis, G. (1999). The natural history of agents. In L. Gulyás, G. Tatai, & J. Váncza (Eds.), *Agents everywhere* (pp. 24–48). Budapest: Springer.
Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. New York: Oxford University Press.
Luhmann, N. (1995). Why systems theory. *Cybernetics & Human Knowing, 3*, 3–10.
Luisi, P. L. (2003). Autopoiesis: a review and a reappraisal. *Naturwissenschaften, 90*, 49–59.
Maes, P. (1995). Artificial life meets entertainment: life like autonomous agents. *Communications of the ACM, 38*, 108–114.
Maturana, H., & Varela, F. (1973). Autopoiesis. The organization of the living. In H. Maturana, & F. Varela (Eds.), *Autopoiesis and cognition. The realization of the living* (pp. 63–135). Dordrecht, The Netherlands: D. Reidel Publishing Company, 1980.
Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Boston: Reidel.
Moreno, A., & Barandiaran, X. (2004). A naturalized account of the inside-outside dichotomy. *Philosophica, 73*, 11–26.
Moreno, A., Umerez, J., & Ibanez, J. (1997). Cognition and life. The autonomy of cognition. *Brain & Cognition, 34*, 107–129.
Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135–183.
Quine, W. V. O. (1969). *Ontological relativity and other essays*. New York: Columbia University Press.
Rocha, L. M. (1996). Eigenbehavior and symbols. *Systems Research, 13*, 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*, 209–228.
Ruiz-Mirazo, K., & Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life, 10*, 235–259.
Ruiz-Mirazo, K., Pereto, J., & Moreno, A. (2004). A universal definition of life: autonomy and open-ended evolution. *Origins of Life and Evolution of the Biosphere, 34*, 323–346.
Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice Hall.
Varela, F. (1979). *Principles of biological autonomy*. New York: Elsevier.
Varela, F. (1992). Autopoiesis and a biology of intentionality. *Proceedings of a workshop on autopoiesis and perception*. Dublin: City University.
Varela, F. & Bourgine, P. (1992). Towards a practice of autonomous systems. In *Proceedings of the first European conference on artificial life*, (pp. xi–xvi).
Varela, F. J., Maturana, H., & Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. *BioSystems, 5*, 187–196.
Vijver, G. (1997). Who is galloping at a narrow path. Conversation with Heinz von Foerster. *Cybernetics and Human Knowing, 4*, 3–15.
von Foerster, H.. (1960). On self-organizing systems and their environments (Reprinted in von Foerster, H. (2003)), (pp. 1–19).
von Foerster, H. (1969). What is memory that it may hindsight and foresight as well? In S. Bogoch (Ed.), *Proceedings of the third international conference: The future of the brain sciences* (pp. 19–64) New York: Plenum Press, (Reprinted in von Foerster, H. (2003), pp. 101–132).
von Foerster, H. (1976)*Objects: Tokens for (eigen-) behaviors, Vol. 8*. ASC Cybernetics Forum. (Reprinted in: von Foerster, H. (2003), pp. 261–71). [Page numbers in the text refer to the reprint]91–96.
von Foerster, H. (1981). *Observing systems*. CA, USA: Intersystems Publications.
von Foerster, H. (2003). *Understanding understanding. Essays on cybernetics and cognition*. New York: Springer-Verlag.
von Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning*. The Falmer Press.
Wooldridge, M., & Jennings, N. R. (1995). Agent theories, architectures, and languages: a survey. In Jennings Wooldridge (Ed.), *Intelligent agents* (pp. 1–22). Berlin: Springer-Verlag.

Please cite this article in press as: Arnellos, A. et al., Towards the naturalization of agency based on an interactivist account of autonomy, New Ideas in Psychology (2009), doi:10.1016/j.newideapsych.2009.09.005