Extending Artificial Agency to Incorporate Ethics-oriented Social Interaction

Argyris Arnellos, Thomas Spyrou and John Darzentas
Department of Product and Systems Design,
University of the Aegean,
Ermoupolis, 84100
Syros, Greece
{arar, tsp, idarz}@aegean.gr

Abstract
AI is still far away from the implementation of a truly autonomous agent, and so far, attempts to build Autonomous Moral Agent (AMAs) have been concentrated on the ad hoc integration of a moral theory to the mechanisms of an Artificial Agent (AA). However, it would be more natural if morals were to evolve within the cognitive realm of the artificial system rather than function separately. A systemic framework of second order cybernetics and self-organization is proposed as the basis for the design and implementation of artificial AMAs. In this proposal information is not taken in the ordinary sense and ethics are reflected in the quality of information processing by the system. This processing is made through the creation of new meaning structures, based on the intentions and goals of the agent. In this perspective, being an autonomous agent is equivalent to being an AMA. The limitations of the framework from an implementation point of view are indicated and an extension of it via the incorporation of semiotic processes is suggested.

1 Introduction
The spread of the notion and use of artificial agency in contemporary information systems raises questions as to whether the former could morally operate within the processes of the latter. The notion of agency consists of at least the notions of autonomy, intentionality, meaning and information, but it seems that their interrelations as well as their activation outside or inside the agent result in different theories of cognition and interaction. Purely computational, connectionist and evolutionary approaches to cognition cannot provide the designers with a framework capable of implementing an artificial Autonomous Moral Agent (AMA), mainly because of the way they handle and use the notion of information. Thus, this paper begins with a definition of the notion of moral agent and its relative concepts. It continues with an analysis of these concepts with respect to the basic theoretical approaches from Cognitive Science and AI to research and analyse cognition and to design Artificial Agents (AAs), as well as with the correspondence of these approaches to the basic and most well known theories of ethics. The approaches are the computational approach, the connectionist, the evolutionary/learning approach and the 2nd order cybernetics approach.
The ways these concepts are encountered in each theoretical approach characterises the design of the AA and consequently the theory of ethics that it is capable of adopting. In each case, the advantages and disadvantages are discerned and form the basis for the use of a systems-theoretical framework which appears capable to support the design of

AMAs, by giving a new significance to the concept of morality and shifting ethics towards self-organisation through interaction.
This framework could help to overcome many of the significant problems of the objective rule-based theories of ethics, by considering the cognitive and moral modules of an autonomous agent as a whole.
As a result, in this framework, being an autonomous agent is equal to being an autonomous moral agent, which gives a new orientation to the design of AMAs.
The last sections of the paper discuss the problems of self-organised systems from an implementation point of view, where the complements of the Peircian sign processes are suggested as a way forward towards the implementation of an artificial AMA, and that, ethics will emerge out of necessity for purposeful interaction of the agent with its environment.

1.1 Moral Actions Need Autonomous Agents

Contemporary problems in Computer Ethics base their attempts for solutions on analysing the relationship between ethics and information**,** [Stahl, 2004], [Floridi, 2004] due to the 'wider' recognition of computers as information processing systems. For the purpose of this paper, ethics are considered as the reflection of the moral actions of an agent towards its environment. Specifically, ethics is the reflection of a code which suggests how an agent should and should not act in order to increase or at least safeguard the welfare of its environment. In this respect, information can be considered as knowledge or facts acquired from the agent's interaction, and used by the agent to base its decisions on how to act. Finally, agency, is defined here in its stronger notion, that is the one usually applied to human agents. Hence, it can be said that the agent is the system which exhibits the following properties:

- interactivity: the ability to perceive and act upon its environment by taking the initiative;
- intentionality: the ability to effect goal-oriented interaction by attributing purposes, beliefs and desires to its actions;
- autonomy: the ability to operate intentionally and interactively based only on its own resources.

With regard to the three basic properties, there appears to be an interesting interdependence between them, in the form of a circular connection between them. As [Collier, 1999] suggests, there is no function without autonomy, no intentionality without function and no meaning without intentionality. The circle closes by considering meaning as a prerequisite for the maintenance of system's autonomy during its interaction. Combining the view that ethics refer to the way an agent should or should not act, with the basic properties of an agent and the fact that an agent's action is based on, or somehow connected with, the concept of information, it can be deduced that a moral agent is an autonomous system whose intentional, therefore meaning-based interactions are driven by information processes. Given this abstract but coherent description of a moral agent, the question that arises is, in the context of Computer Ethics, whether it is possible for an AA to morally interact with its environment.

2 Approaches to the Design of Artificial Agents

2.1 Computational Artificial Agents

The first attempt in creating an autonomous AA takes a computational approach, which is based on the hypothesis that the mind is supposed to process symbols that are related together to form abstract representations of the environment. Computational-based AAs are purely formal systems and their symbols are related to an apriori correspondence with externally imposed meaning [Newell, 1980]. They are processing information based on a static meaning structure, which cannot be internally changed and grounded in order to adapt to the continuously changing demands of a dynamic environment. Since these systems by their nature, separate syntax and semantics, and manipulate their externally given representations as sequences of symbols being manipulated by also externally given rules, they will not be able to produce inherent meaning in order to intentionally classify their environment.
There is no need for self-organization of the system, and all its variety is externally selected. Additionally, these systems are characterized by a high degree of causality which, by means of computationalism, supports the view that all intentional content is a kind of information, which is externally transmitted by a causal flow [Smith, 1999].

Since purely computational-based artificial systems lack the property of intentionality because intentionality has its source outside the artificial system, primarily inside its designer, consequently, the system exhibits no autonomy and functionality other than that of its designer. In addition, since all its functionality comes into it in the form of rules given during its design phase, its functionality and therefore its interaction with other systems is based on a predetermined and universal information sets. In such systems, the concept of information is clearly translated into the "data+meaning" model which assumes that two agents are interacting with each other by exchanging objective meaning, in terms of an information structure whose syntax will produce in them the same semantics wherever and whenever they are processed.

From this analysis follows that a purely computational, symbol-based artificial system cannot meet the criteria given above for a moral agent. However, it would be interesting to see if some of the information-oriented theories of ethics can be suited (due to their explicit and rule-based nature) to a purely symbolic artificial system. A similar correspondence between certain moral theories and well-known implementation architectures, with a strong focus on the sufficiency of computing capabilities of the artificial system, has been undertaken in [Allen, Varner and Zinser, 2000]. The examination attempted in the present paper intends to demonstrate the inseparability between cognitive and moral capabilities from a theoretical, as well as from an implementation point of view.

Computationalism and Utilitarianism

The top-down nature and the explicit and predetermined functionality of purely computational artificial systems make them a good candidate for the implementation of a consequentialist moral theory [Anscombe, 2002], the utilitarianism [Mill and Sher, 2002], according to which an agent's act is considered ethically correct if it contributes to an aggregate state of maximal happiness and utility. In this view, an immoral action is the one which leads to a state of minimized utility and unhappiness. Although utilitarianism has not the genuine characteristics of a computational-based moral theory it seems that it can

be regarded as such. First of all, it is not a teleological theory, as it uses normative rules in order to derive the moral action that should be followed (this is decided based on the criterion of maximum utility and by using a consequentialist approach in order to choose among several probable actions) [Anscombe, 2002]. Such an approach strongly rejects the notion of inherent intentionality of a moral agent, as it is not acting based on what its autonomous and history-based functionality suggests that it should, but based on what the externally given rule of preservation of maximum utility asserts. In this way, it resembles a meaningless system whose only concern is to compute the concrete utilities of a group of probable actions, which, from an implementation-oriented point of view leads to many problems. As has been noted in [Allen et al., 2000], the assignment of numerical values to each effect of each action would produce problems of computational intractability. On the other hand, even if such problems were solved by supercomputers there would be a much larger problem to confront. Namely, the algorithms which would attribute the numeric values to each action would need to be decided upon and this is not a trivial problem. Claiming awareness of the total causal chain of all possible actions (at least for a specific framework) is not always realisable and could indeed be dangerous. Such a thought would support the existence and use of purely objective information of a mechanistic nature [Brier, 1992], which, has long been bypassed by the theories of non-linear dynamic systems [Prigogine, 1997]. At this point, the utilitarian designer may choose to use a dynamic non-linear algorithm for the ascription of the numerical values, but this will not solve the problem completely either. Non-linearity can supply no concrete values, so the system will have again to choose among the set of values of the chosen attractor and this could lead to an infinite loop. Even if the infinite depth of such calculations became finite by the establishment of a horizon of the action after which no further utility-based assessment can be claimed [Allen et al., 2000], the problems remain. This technique does not coincide with the essence of the utilitarian theory of ethics and moreover, the deeper problem of the inherent intentionality persists. Particularly, as will be discussed below, the use of non-linear equations in the mechanism of cognition requires a totally different way of dealing with the concept of information [Kelso, 1997].

Kantian moral theory and rule-based systems

Another rule-based moral theory is Kant's theory of ethics, which is categorised as a deontological moral theory and is traditionally contrasted with utilitarianism. Kant introduced the categorical imperative, to which all maxims (i.e. subjective rules of action) must conform in order for an action to qualify as being moral. The centre of Kant's theory is the subject but not its intentionality, as each subject should act freely and out of respect for the moral law rather than out of its inclination to fulfill its own desire for happiness. Thus, as stated in [Beck, 1989] the first criterion for the categorical imperative is that a subject's act should only take place according to that maxim by which the subject can, and at the same time desire, that it should become a universal law. The second requires the subject of the act to respect all other subjects and the third requires that the maxim must be autonomous, which means that the subject itself should indicate and decide the maxim and the laws to be followed for its fulfilment.

The problems regarding Kant's moral theory and its implementation by a purely computational artificial system are deeper than those in the utilitarian situation. On first sight, as [Allen et al., 2000] have noted, an artificial system could be programmed with specific cognitive processes which will be an integral part of the agent's decision-making procedure. According to [Stahl, 2004] this could be counted as a Kantian moral computing system and in combination to the logic nature of the notion of universability makes

computing systems a hopeful candidate for the implementation of the theory. On reflection, however, there are some fundamental problems. As noted above, ethics refer to the way an agent should or should not act, hence, the decision-making procedure that will guide the agent's action should be the one which will guide its total action. It cannot be any other way, especially in the Kantian theory. Information, which is the base for the system's decision-making procedures cannot be divided and ascribed separately to each functional module. From an implementation point of view, this requires a purely rule-base computational system, which carries all the problems discussed before, thus it cannot count for as artificial AMA.

In reply, a Kantian advocate might suggest that due to the third criterion of autonomous maxim mentioned above it would not be correct to implement the Kantian theory in a purely computational artificial system. Instead, the functionality of the categorical imperative would require an autonomous agent which should be deciding its actions based on the maxims, which in turn are in accordance with its meanings. Since, computational-based systems do not have such properties, they are not good candidates for the implementation of the theory. So, it seems that Kant's universalism and categorical imperative needs an autonomous AA, in which the designer responsible for the procedures of the moral action would integrate the respective functionality. Since there is no truly autonomous AAs, in the sense defined above, at the moment this cannot be achieved. But, even if there were such artificial systems, it can be speculated that the introduction of such a procedure would interfere with the other functional procedures. The functionality of the 'how to act' procedure would either be the dominant one, or it would not be counted at all by the system itself. What can be concluded is that in general, a moral theory with such an abstract rule (the rule of acting according to the categorical imperative) cannot be given from above, but it should emerge within the more general functionality of the agent. Such bottom-up self-organizing architectures may be more appropriate for the emergence of an autonomous AA, as it will be discussed below, but it is not at all certain whether such a system will be able to develop a Kantian moral theory.

The Moral Turing Test cannot pass the Chinese Room Argument

As has been mentioned previously, purely computational systems assumes an objective world view and a totally externalized notion of information and meaning. Although this is enough to support the functioning of an artificial agent with high computing capabilities but with externally imposed moral procedures, it is in no way adequate for the design and implementation of an artificial AMA. However, in [Allen et al., 2000] it is suggested that given the disagreements about ethical standards as well as what constitutes a cognitive system which in turn is able to act morally, a solution to the problem of defining an artificial AMA will be to apply the Turing test to conversation about morality. In the case where an observer of a discussion between a human and an AA cannot identify the machine at above chance accuracy, then the machine can be considered as a moral agent. This moral Turing test (MTT) is a suitable solution from a solipsistic perspective. Searle's Chinese Room Argument [Searle, 1990] explains why such a system can behave like a moral agent but would never be a moral agent in itself.

Moral reasoning needs first of all basic reasoning processes, and meaning is their primary ingredient, which plays a serious role in the design of an artificial AMA. Thus computational systems do not appear to be a good candidate for the implementation of such agents due to all problems discussed above. More promising might be connectionist

architectures that emphasise learning, but they also present serious limitations as analysed in the next section.

2.2 Connectionist Artificial Agents

There has been an attempt to confront the apparent lack of symbol grounding in purely computational systems by the introduction of connectionist architectures. Connectionism argues that the mind is a network of interconnected neurons that gives rise to a dynamic behaviour that can be interpreted as rules at a higher level of description. Here, the dominant view is that mental elements are a vector distribution of properties in dynamic networks of neurons and the proposed solution for a proper modeling of the thinking process is the set-up of parallel distributed architectures [Smolensky, 1988].

The basic and most important advancement of connectionist architectures over those used in purely symbolic artificial systems is in the constitution of a system's representations. In connectionist architectures, representations are massively distributed, being stored as weights between neurons. There is not a one-to-one correspondence between individual neurons and individual representations. This is the reason that such architectures are called subsymbolic, in contrast to symbolic architectures, where representations are mapped to symbol tokens in order for rules to operate over them, thus making them purely symbolic. Therefore, due to the parallel and distributed nature of their representations connectionist architectures may sometimes bear richer syntactic structures. However, despite this, [Fodor and Psyslyn, 1988] among many others argue that the form of the computation, whether logico-syntactic or connectionist, is merely a matter of implementation, and in addition, the implementation of computation, whether classical or connectionist, lies in causal processes. The only change is the introduction of a more sophisticated technique for the correspondence between symbolic input and its weight-based processing. Information remains as something which is completely external and given to the system and there is no true autonomy as an artificial neural network is able to change the weights of a particular transfer function but it is not capable of altering the transfer function itself. In this perspective and in relation to intrinsic creation of meaning, connectionist architectures cannot offer a significant difference to the design of autonomous AAs, or artificial AMAs.

2.3 Evolutionary and Learning Approaches to Artificial Moral Agents

An alternative approach to symbolic and subsymbolic approaches to autonomous AAs is behaviour-based architectures that follow a bottom-up approach. Here, autonomy is modelled on a biological system's capacity to interact with its environment, rather to represent it internally [Ziemke, 1998]. Additionally, it also studies the intelligent behavior as a result of adaptation at the cognitive and social level [Dautenhahn, 1995]. The main idea is to start with the design of simple modules with multiple interaction capabilities, while expecting their interaction to give rise to complex adaptive behavior [Brooks, 1991]. These kinds of architectures are based on the concept of intelligence not as formal and abstract input-output mapping, but as a property arising from the system's physical interaction with their environment. Some interesting cases in this direction are the design of evolutionary connectionist architectures. In the case of AMAs these evolutionary techniques are inspired by the mechanisms of natural selection and allow for the evolution of a large number of individual behaviours, each capable of representing a different 'morality'.

Although this approach avoids the use of an explicit and abstract moral theory by allowing each person responsible for instructing each artificial system to do so based on his/her own

morality, there are still some fundamental problems which prohibit such systems to be considered as autonomous AMAs.

To begin with, teaching may solve the problem of having pre-defined the content of all of an agent's interaction but it does not solve the problem of inherent meaning. The variety of the interaction is externally imposed, as each module's behaviour is pre-programmed in an algorithmic manner. The nature of such an agent remains computational and carries with it all the respective problems. Additionally, the functionality of the interaction based on natural selection implies the existence of purely external information structures corresponding to internal representations. The selection of the latter is based not only on the pre-decided mental functioning but also on their relation to the environment and the historical processes through which they were selected for a particular moral action. The latter cannot be of use in the absence of inherent meaning. This will cause adaptation problems in the case of AA's interaction with an agent of different moral values. The only thing it could do when confronted with a new situation, it is to ask for recommendation from the instructor, which of course could have 'immoral' judgments for the particular situation. But since it is able to learn, one could assume that it could be instructed on how to act for the respective situation from another instructor, that is, the one it interacts with. In that case, it would have no use at all, since it is supposed that an autonomous AA should be able to help other agents by its actions. One can go even further and assume that AAs have by their nature other properties (i.e. extreme computational and arithmetic capabilities), so, it is quite permissible to have ad hoc provisional instruction on ethics which, then, would provide the moral part of their actions. The problem is this, however: an AA with no inherent meaning could never judge if and when it is appropriate to change instructor. In the case where it is just programmed to take moral lessons from each agent which it interacts, it is not autonomous and no adaptation can be expected from it.

Evolutionary connectionist approaches introduced new concepts in the design of AAs. The top-down approach and the purely symbolic nature of the representations in purely computational systems is replaced by the bottom-up co-operative learning based on external selection. Hence, the role of the environment in the interaction becomes stronger and the artificial system acquires some kind of internal memory which is in accordance with results of the respective historical processes made during the system's evolution. Despite the absence of merely symbolic mapping, the system engages in interaction using a more dynamic and situated mechanism for the mapping of input data to output actions. This mechanism is responsible for the merely causal anchorage to external information structures of the internal system's representations. Hence, natural selection leaves the environment to choose the results of system's purely behavioral actions, therefore, the semantic part of the interaction resides in the environment and not in the system.

It would appear that so far, in spite of whether a system is designed top-down or bottom-up, its cognitive capabilities are based on the ways the designers have decided to connect, or map meaning to, the system's internal states with its environment. This on its own is not sufficient, since an AMA should first of all be autonomous and the requirements for autonomy need the formation of an agent's internal meaning structures, as well as the integration of system and environment as a whole. The problem of intrinsically generated meaning in an artificial system requires a holistic and systemic approach, which will complement the properties of externally selected evolutionary systems, while shifting the semantics inside the AA. In answer to this, the systems theoretical framework of second

(2nd) order cybernetics and self-organisation seems to offer such an approach. A description of the important characteristics of 2nd order cybernetics and self-organized systems is given in the next section. Subsequently ethics are analysed in this systems theoretic framework, which provides a new perspective regarding moral acting. The attempts to incorporate this new approach to the development of autonomous AMAs is discussed.

3 Second order Cybernetics and Self-organisation

Cybernetics has from its beginning been interested in the similarities and differences between autonomous, living systems and machines. While in purely mechanistic science the system's properties are distinguished from those of their models, which depend on their designers, in 2nd order cybernetics the system is recognised as an agent in its own right, interacting with another agent, the observer, which is in turn a cybernetic system [Heylighen and Joslyn, 2001]. To understand this process, where the cybernetician enters his own domain and has to account for his own activity, a theory is needed wherein cybernetics becomes cybernetics of cybernetics, or 2nd order cybernetics [von Foerster, 1995]. Second order cybernetics made a very successful attempt to clearly distinguish itself from the pure mechanistic approaches by emphasizing autonomy, self-organization, cognition and the role of the observer in modelling a system.

The point of departure of 2nd order cybernetics is the actions of distinction and observation. For a system to exist in its own right, it must be defined or delimited, creating a boundary between itself and the environment. It can achieve this by observing its boundary and this is a prerequisite for it to become a 'system'. As the system is able to observe the distinctions it makes, it is able to refer back to itself the result of its actions. This is the phenomenon of self-reference, which gives the ability to the system to create new distinctions (actions) based on previous ones, judging its distinctions and increasing its complexity by creating new meanings in order to interact [Luhmann, 1995]. This self-referential loop dismisses the classical system-environment model, according to which the adaptation of a system to its environment is controlled externally and according to the course of a learning process and is replaced by a model of systemic closure. Due to system closure, environmental complexity is based solely on system observations, thus, system reality is observation-based. In contrast the self-reference of an observation creates meaning inside the system, which is used as a base for further observations in order to reduce external complexity. The system which operates on meaning activates only internal functions and structures, which von Foerster calls eigenvalues [von Foerster, 1984] and which serve as points of departure for further operations during its interaction with the environment. Indeed, this closure is operational in so far as the effects produced by the system are the causes for the maintenance of systemic equilibrium by forming new more complex organizations. Thus, each new operation based on observations is a construction, it is an increase of the organisation and cognitive complexity of the system. von Foerster was among the first to attempt to describe this process of the phenomenon of self-organization: an increment of order [von Foerster, 1960].

Accordingly, a self-organized system establishes and changes its own operations, in this sense being 'autonomous', while at the same time, being dependent of an environment which as [Ashby 1962] says, pre-programmes the points of relations for the self-conditioning. Therefore, self-reference can only exist in relation to an environment. In

fact, the 'order from noise' principle suggests this very idea, according to which, a self-organized system can increase its order by moving to a higher level of organization through the selection of environmental perturbations and their subsequent incorporation into the structure of the system [von Foerster, 1984]. But due to the nature of systemic closure, all system's selections are internally produced and moreover, they are selected by a totally internally produced area of distinctions. The environment cannot contribute with anything, since it contains no information. [Luhmann, 1995]. The way such a system interacts with its environment is discussed in the next section, looking at the role of information in 2nd order cybernetic systems.

3.1 Information in Second Order Cybernetics

The proponents of 2nd order cybernetics consider that a self-organised system does not interact with its environment by responding to externally given objective information. Due to the conditions posed by operational closure, information is something that is created inside the system itself, as an internal regulation of its organisation. Moreover, the system must itself produce that which is information for it, in order to establish those structures which are considered as knowledge for itself [Luhmann, 1995].
Maturana and Varela in an attempt to form the basis and define biological systems in this framework, introduced the concept of autopoiesis [Maturana and Varela, 1980], which transfers the principle of self-reference from the structural to the operational level [Luhmann, 1995]. Now, the role of the environment is constrained to that of an irritation to which the system would adapt using only its own resources and keeping its operational closure. According to the theory of autopoiesis, there is no information at all. It can only be socially ascribed to a process of interaction between two systems from other observers. Hence, there is no representation of the environment, but only the system's own constructions. In this perspective, the interaction between an autopoietic system and its environment takes place in terms of a structural coupling between these two. It is implied that autopoietic systems do not need external information in order to self-organise. The structural coupling makes them self-organize in order to compensate for the perturbations.

3.2 Ethics in the Framework of Second order Cybernetic

In the discussed framework, communication cannot be defined as a transfer of information from one place to another. It is rather a deformation, possibly caused by events in the environment, which each system compensates according to its own self-organisation. Communication and interaction then between two agents in 2nd order cybernetics is considered as a double structural coupling between two closed systems, where each is internally creating information [von Foerster 1993**]**. Therefore, it would be difficult to locate and identify the moral part of such agents and even more difficult to explain the role ethics can play in the interaction between such agents.

According to [von Foertster, 1984] the origins of ethics is where one cognitive system computes its own computations through those of the other. This statement may seem somewhat strange and complicated at first glance, but is in fact a descriptive and accurate statement in regard to the notion of ethics in 2nd order cybernetics. In the framework discussed, morality is a way of observing, using the distinction between right and wrong. Since a 2nd order agent operates only on its own distinctions, morals cannot be given an objective or rational basis. Morals are based on each agent's choices during its history of interactions with its environment. Each agent maintains its autonomy and its moral code. As Thyssen [Thyssen, 1992] says, the question it raises is if, since morals cannot be

justified objectively and furthermore, no moral agent can prove its superiority, it is possible to develop rules which apply to the relation between agents and which, therefore, cannot just be based on the rules of any one agent. He argues that such rules should be shared rules, which are defined as ethics. Therefore, ethics is the shared rules of morals, it is the morals of morals, or even better, it is second order morality.

From the point of view of 2nd order cybernetics, the morality of one agent is external to the other. When two agents purposefully interact one cannot affect the other, since both of them are the dominants of their closure. But, each one is an environment of the other, which can be observed and co-related if they both agree on some shared values. This is the reason why [von Foerster, 1984] sees ethics as a self-adaptation, as it is founded in between two cognitive systems. When two agents interact, they will adapt to each other in some fashion. The means of their adaptation are private to each one and they are related to their variety, to the richness and complexity of their self-organisations. Ethics is not developed inside the agent but between them, each having its own morals. It arises in the interaction during their self-adaptation, but only if they succeed in defining values and some key interpretations which they can all accept [Thyssen, 1992].
Even in the absence of shared values, where a conflicting situation is implied and the forms of meaning of each agent cannot be met, their purposeful interaction may result to the emergence of ethics. It is the ethical process that matters and not each agent. The latter will be affected and accept the outcome of the interaction if she has been involved in it, that is, if it has participated in the mutual structural coupling. Considering that the agent's tools for interaction are hidden in the forms of meaning which it carries through its self-organisation, and which deforms in order to compensate for the uncertainty of the environment, ethics result as a consequence of the uncertainty. Therefore, ethics are not ideals, but an emergent necessity.

Finally, if ethics are seen as shared values that emerged during interaction, they cannot be universal, but they are by nature contextually dependent. Additionally, since each agent's adaptation is in accordance with its variety, which share an asymmetrical relation to the other's, ethics cannot express the values of each agent in its totality. Hence, they are not defined once and for all. Ethics in this sense must be seen in an evolutionary perspective, so that with their introduction they create a base of mutual expectations and they introduce new criteria, on top of which new interactions between more ethically developed agent's can take place. This is what is meant by [von Foerster 1990] when he says that ethical action is the one which increases the number of choices to each one of the interacting participants.
Considering the theoretical framework of 2nd order cybernetics and the incorporated notion of ethics, the design of an AMA proposed is based on the design of an artificial self-organised agent. The next section discusses the problems with contemporary approaches to the design and implementation of artificial self-organised systems and proposes a solution that offers the enhancement of their interactive capabilities.

4 Designing Artificial Self-organized Moral Agents

Contemporary AI is still far away from implementing a 2nd order cybernetic agent [Groß and McMullin, 2001]. The most important attempt to computationally mimic the nature of a self-organised system has been made in the context of dynamic systems theory. In this context, which has as a central point the system's nonlinear dynamical processing, the brain is seen as a dynamical system whose behavior is determined by its attractor

landscape [Port and van Gelder, 1995]. Dynamic cognitive systems are self-organised by a global co-operation of their structure, reaching an attractor state which can be used as a classifier for their environment. In such a case, symbolic representation disappears completely and the productive power is embodied within the network structure, as a result of its particular history [Beer, 2000].
In computational approaches to self-organisation meaning is not a function of any particular symbols, nor can it be localised to particular parts of the self-organised system. Their ability for classification is dependent only on the richness of their attractors, which are used to represent, - though not in a symbolic way,- events in their environments. Therefore, and here lies the problem, their meaning evolving threshold cannot transcend their attractor's landscape complexity, hence, they cannot provide architectures supporting open-ended meaning-based evolution.
At the theoretical level and especially in the framework of 2nd order cybernetics, this problem is concentrated on defining the means by which the structural coupling will take place. As mentioned before, 2nd order cybernetic systems admit no functional usefulness to representations and they regard information as something merely internal. On the other hand, many proponents of the dynamical approach find representations a necessary property in order for the system to exhibit high-level intelligence [Clark and Eliasmith, 2002], or even any kind of intentional behaviour [Bickhard, 1998], as long as representations emerge from the interaction of the system in a specific context of activity. Consequently, the incorporation of a process to support the vehicle of the representation which carries internal information about an external state seems imperative [Brier, 2001]. This process would provide the appropriate interactive dimension to the self-organising system. It would comprise the appropriate mechanisms to support and guide a system's interaction with the environment, formed by other systems.

Semiosis can be seen as such process which will drive the system into meaningful open-ended interaction. In [Arnellos, Spyrou and Darzentas, 2003] the process of semiosis and especially Peircian triadic semiosis [Peirce, 1998] are presented in some detail as a proper mechanism in order to complement the interaction of 2nd order cybernetic systems in a dynamic information environment, as well as, the ability of such processes to model intentional interactions. The suggested framework proposes a way out of the poor classification capabilities of the artificial dynamic systems. Although such systems seem to exhibit a (not very high) degree of operational closure, which is, by all means, a prerequisite for the implementation of 2nd order cybernetic agents, they also exhibit informational closure, fact which constrains them while they interact with their environments. As it is mentioned above and as it can be easily implied from the framework of 2nd order cybernetics, for an AA to be autonomous, it should be able to internally extend its structural representations in order to dynamically classify its environment. Semiotics seem to help in this way by supplying the designer of an AA the tools to model the interactive part of the whole self-organised process. Hence, they can work as a complement which gives the proper input to the purely dynamical and self-organised part, while their more significant contribution is that they are essentially driving the artificial systems self-organisation. In addition, Peircian semiotic processes support the incorporation of pragmatic meaning [Collier, 1999b] into the system, a property which is a strong requirement for the design of truly autonomous AAs.

5 Summary and Conclusions

Throughout history many attempts have been made to define morals on an objective basis. As [Thyssen, 1992] argues, the reason for objectivity is that it was allegedly implying obligation, while a relative approach will produce a balanced situation between good and bad. As it was stressed in this paper, such theoretical approaches to morality cannot be imposed as an extra moral module in an AA's cognitive system. Also, from an architectural point of view, they cannot be used as the basis for designing a moral agent, as their functionality will not result in the artificial system's autonomy. The same argument holds for learning approaches to morality. An agent which is instructed how to act morally, cannot be said to have autonomous-based morality.

Purely cognitivist/connectionist and evolutionary and learning frameworks have been examined and shown inappropriate for creating artificial AMAs. The reason for this is that they make use of a purely objectified and mechanistic notion of information, which originates from outside the system, and must be inserted into it. However, morality is a product of evolutionary co-adaptation of shared expectations, not a product of rational design or learning processes. The systemic framework of 2nd order cybernetics provides the passage from a mind-less morality [Floridi, 2001] to a mind-oriented morality in the artificial domain. Self-organisation immerses morals into the cognitive capabilities of the agent and imputes to them a purely subjective nature. In that case, information is not something that is inserted into the system from the outside, but is internally and subjectively created. The system now interacts with its own resources and its own moral values. The environment provides only those perturbations which have their own morals striving for adaptation with the ones of the system. Ethics, as second order, comes as a result of this interaction and are incorporated into each agent's self-organisation in order to be used as a base for further, richer interactions. In this perspective, self-organised architectures do not have to take care of concrete moral modules of an artificial AMA, since these do not exist. Rather the self organised architectures can better reorient themselves in finding ways to complement and drive the system's self-organisation into a purposeful interaction. The richness and nature of the Peircian semiotic process appears to offer a useful toolset for this direction. Ethics will then emerge as a necessity for adaptive interaction.

References

Allen, C., Varner, G., Zinser, J. (2000), Prolegomena to Any Future Artificial Moral Agent, Journal of Experimental and Theoretical Artificial Intelligence 12, 251–261.

Anscombe, G. (2002), Ethics, Religion and Politics: The Collected Philosophical Paper, Blackwell Publishers.

Arnellos, A., Spyrou, T. and Darzentas, J. (2003), Towards a Framework that Models the Emergence of Meaning Structures in Purposeful Communication Environments, The 47th Annual Conf. of the Int. Society for the Systems Sciences (ISSS) 3(103).

Ashby, W. R. (1962), Principles of the Self-organizing System, in Foerster, Heinz von. and Zopf, W. G. (ed.), Principles of Self-organization, New York.

Beck, L. (1989), Immanuel Kant: Foundations of the Metaphysics of Morals, Prentice Hall.

Beer, R. D. (2000), Dynamical Approaches to Cognitive Science, Trends in Cognitive Science, 4(3) 91-99.

Bickhard, M. (1998), Robots and Representations in (R. Pfeifer, B. Blumberg, J. -A. Meyer, & S. W. Wilson, (ed), From Animals to Animates 5, MIT Press, Cambridge, MA.

Brier, S. (1992), Information and Consciousness: A Critique of the Mechanistic Concept of Information, Cybernetics & Human Knowing, 1, 2/3, Aalborg.

Brier, S. (2001), Cybersemiotics: A Reconceptualization of the Foundation for Information Science, Systems Research and Behavioral Science. 18(5), 421-427.
Brooks, R. (1991), Intelligence Without Representation. Artificial Intelligence. 47, 139–159.
Checkland, P. and Holwell, S. (1998), Information, systems and information systems: making sense of the field, John Wiley & Sons.
Clark, A. and Eliasmith C. (2002), Philosophical issues in brain theory and connectionism, In Arbib, M. (ed), Handbook of Brain Theory and Neural Networks, MIT Press, Cambridge, MA.
Collier, J. (1999), Autonomy in Anticipatory Systems: Significance for Functionality, Intentionality and Meaning in Dubois, D. M. (eds.) The 2nd Int. Conf. on Computing Anticipatory Systems. Springer-Verlag, New York.
Collier, J. (1999b). The Dynamical Basis of Information and the Origins of Semiosis. in Taborsky, E. (ed), Semiosis. Evolution. Energy Towards a Reconceptualization of the Sign, 3, 111-136.
Dautenhahn, K. (1995), Getting to Know Each Other-Artificial Social Intelligence for Autonomous Robots, Robotics and Autonomous Systems. 16 (2-4), 333-356.
Floridi, L. and Sanders J.W. (2001), On the Morality of Artificial Agents, in Introna, L. and Marturano, A. (ed), Proceedings Computer Ethics: Philosophical Enquiry – IT and the Body, Lancaster.
Floridi, L. (2004), Information Ethics: On the Philosophical Foundation of Computer Ethics, ETHICOMP Journal (www.ccsr.cse.dmu.ac.uk/journal) Issue - Vol. 1 No. 1.
Fodor, J. A. and Pylyshyn, Z. (1988), Connectionism and Cognitive Architecture: A Critical Analysis, Cognition, 28, 3-71.
Foerster, Heinz von (1960), On self-organizing systems and their environments, in Yvotis, M. and Cameron, S. (ed.), London, Pergamon Press.
Foerster, Heinz von (1984), Observing Systems, Intersystems Publications, CA, USA.
Foerster, Heinz von (1984), Disorder/order: Discovery or invention? in Livingston, P. (ed), Disorder and order, Saratoga CA.
Foerster, Heinz von (1990), Ethics and second-order cybernetics, Cybernetics & Human Knowing, 1(1).
Foerster, Heinz von (1993), For Niklas Luhmann: How Recursive is Communication? Teoria Sociologica 1(2), 61-88.
Foerster, Heinz von. (1995), The Cybernetics of Cybernetics (2nd edition), FutureSystems Inc., Minneapoli.
Groß, D. and McMullin, B. (2001), Towards the Implementation of Evolving Autopoietic Artificial Agents, in Kelemen, J. and Sosík, P. (ed), Advances in Artificial Life, Proceedings of the 6th European Conference. Springer.
Hartshorne, C., Weiss, P. and Burks, A. (ed) (1998), Collected Papers of C.S.Peirce, Thoemmes Pr.
Heylighen, F. and Joslyn C. (2001), Cybernetics and 2nd order Cybernetics, in Meyers, R. A. (ed.), Encyclopedia of Physical Science & Technology (3rd ed.), Academic Press, New York.
Kelso, J. (1997), Dynamic Patterns: the Self-organization of Brain and Behavior, MIT Press.
Luhmann, N. (1995), Why “Systems Theory”, Cybernetics & Human Knowing, 3(2), 3-10.
Maturana, H. R. and Varela, F. J. (1980), Autopoiesis and Cognition: The Realization of the Living, Reidel, Boston.
Mill, J. and Sher, G. (ed) (2002), Utilitarianism, Hackett Pub Co.
Newell, A. (1980), Physical Symbol Systems, Cognitive Science, 4, 135-183.
Port, R. and van Gelder, T. (ed), (1995), Mind as Motion: Explorations in the Dynamics of Cognition, MIT Press, Cambridge, MA.
Prigogine, Y. (1997), The End of Certainty, Free Press.
Searle, J. R. (1990), Is the brain a digital computer?, Proceedings and Addresses of the American Philosophical Association, 64, 21-37.
Smith, W. D. (1999), Intentionality Naturalized? in Petitot, J., Varela, F., Pachoud, B., Roy, J-M. (ed), Naturalizing Phenomenology, Issues in Contemporary Phenomenology and Cognitive Science Stanford University Press, Stanford CA.

Smolensky, P. (1988), On the proper treatment of connectionism, Behavioral and Brain Sciences, 11, 1–74.
Stahl, B.C. (2004), Information, Ethics, and Computers: The Problem of Autonomous Moral Agents, Minds and Machines, 14: 67–83.
Thyssen, O. (1992), Ethics as Second Order Morality, Cybernetics & Human Knowing, 1(1).
Ziemke, T. (1998), Adaptive Behavior in Autonomous Agents, PRESENCE, 7(6), 564-587.