CONCEPT FORMATION: A COMPLEX ADAPTATIVE APPROACH

Mukesh J. PATEL *

ABSTRACT

Concept formation is complex cognitive phenomenon which has been only partially modelled in Cognitive Psychology and AI. Following a detailed and critical evaluation of such models we conclude that their main shortcoming of not being able to explain the nature of the semantics of concepts is because they fail to take into account the role of learning in concept formation. As a radical alternative it is proposed that a more (semantically) complete model would necessarily have to give an account of how concepts are formed as a result of agent-environment interaction, that is mediated by an agents action in its environment and feedback from it. Such a shift in focus renders the investigation of the nature of concept formation as a complex and dynamic adaptive system. In accordance with this shift in perspective we propose and describe a novel methodological approach and a computational model that would support simple concept formation in an autonomous agent, which enables us to investigate concept formation in a more comprehensive manner. Overall we provide a compelling justification of the efficacy of such an ecological approach to the study of concept formation.

Keywords: Learning and development of concept formation, internal/mental representation, autonomous systems and agents, induction and Learning Classifier Systems.

1. Introduction

A novel approach to the study of concept formation will be described here. It is aimed at overcoming limitations of one sort of models of concept formation in Psychology and Artificial Intelligence (AI) that do not give an adequate account of the process of learning and assimilating new concepts. Recent developments in the field of Complex Adaptive Systems in general and Artificial Life in particular provides a better methodology for investigating the process of simple concept formation. More specifically, the aim is to exploit powerful learning (search) techniques such as genetic algorithms (GA's) and parallel distributed processing (PDP) in order to investigate concept formation in an artificial agent or

THEORIA - Segunda Epoca - Vol. IX
1994, Nº 20, 89-108

organism[1]. This is essentially a bottom-up approach, free of any preconceptions about the nature and structure of concepts. Interaction between organism and environment forms the basis for concept formation and their semantic values. Learning about the world is not determined by preconceived conceptualisation, as is often the case in knowledge representation in AI. Instead, we assume that an agent (or artificial organism) learns and forms concepts about its environment[2] by performing goal directed behaviour, such as survival and improved fitness. These sorts of parameters have an important affect on how concepts are formed and what they represent. In other words, the approach enables the investigation of concept formation as emergent properties of the dynamic process of interaction between an agent and its environment motivated by a set of goals that typically ensure the agent's continued survival in an environment. Before describing our approach in more detail, the classical notion of concepts and models inspired by it will be described and evaluated in the next two sections.

2. Concepts: Review and Evaluation

2.1. Classical Notions of Concepts

Concepts are the representational equivalent of words or phrases referring to objects or events in the world (lexical items, is the term normally used in linguistics). Concepts classify objects and events into sets. Each concept is normally assumed to be defined by a set of features. These features correspond to various properties of objects and events (in the world). For example, the concept of a banana would have a feature to denote colour with could be assigned the value, "yellow". How yellow came to mean what it does is not addressed. Instead, most accounts concentrate on the role of features (as symbolic representations) in concept formation and the relationship between concepts (Keil 1989). Since features represent invariant properties across objects and events, their semantic value is assumed to be determined by this correlation[3]. From an information processing point of view concepts are internal representations of information gathered (via suitable transducers) from the world. The semantic value (or intension) of a concept is determined by its defining feature set. Individual features can be members of more than one set. The correspondence between features (sometimes also referred to as, primitive features, in this context) and invariant properties in the world is based on two basic assumptions. First, that the world contains invariant phenomena that can be extensions (referents) of features (and even some basic concepts); features are symbolic encodings which represent specific properties of the world. Second, that a finite set of features corresponding to a set of properties in the world is sufficient to account for all subsequent concept formations and inferences based on them. This assumption has an important implication for modelling emergent phenomena. In order to be formally sound the number and semantic value of basic features has to remain fixed, and therefore have compositional semantics.

There are a number of different Psychological and AI models of concepts (see Wrobel 1990), some of which will be briefly outlined here in order to highlight their basic limitations which our approach is designed to overcome. In so doing, we are aware that, in spite of their limitations, some psychological models account for a large portion of empirical data (see for example, Rosch and Lloyd 1978, and Keil 1989), and that certain AI implementations perform specific task with a fair degree of success (Feldman 1986; Hinton 1986, 1990). However, none of these provide an account of how semantic values are assigned during concept formation. This is because they ignore the role of interaction between agent and its environment in the process of concept formation, which incidentally, also helps ground the semantics assigned to basic features (Hanard, 1990).

2.2. AI (classical) and Psychological Models of Concepts

The Classical Model is one of the earliest and the most basic, rigid and limited in scope. It assumes clear boundaries between concepts and a hierarchical structure to represent relationships between concepts. The nature and function of features determine the clarity of boundaries between concepts. This formal account of concepts, perhaps good model of human competence, fails to predict a large portion of psychological data. For example, there is evidence to suggests that concepts are not very clearly defined in practise (Smith and Medin 1981), and that some concepts seem to have a more central (or basic) role than others (Rosch *et al* 1976) thus upsetting the notion of a hierarchical representational structure. This class of concepts are regarded as equivalent of *natural* categories though that has not proved particularly insightful as far as our understanding of the nature and role of concepts in cognition is concerned. Further, instantiated members of a concept are not considered to be equal; some members are regarded as more typical than others (Rosch and Mervis 1975).

Alternative models to account for psychological data incompatible with the predictions of the classical model have been proposed. A probabilistic based model assigns different saliency (or weight) to features between and within sets defining concepts. This, coupled with a threshold based decision procedure, gives a more realistic account of the role of concept in cognitive processes and (human) behaviour. Exemplar based models are an alternative attempt at overcoming the limitations of the basic classical model (Rosch 1975). These models predict that concepts are defined by exemplars which if properly defined can determine gradations in, and typicality of, membership (Holyoak and Glass 1975; Medin and Smith 1984). Subsequent work in this field has concentrated on providing better accounts of how concepts are used in cognitive processing. This includes designing a better, that is, more reliable and flexible membership decision procedure (see for example, Medin and Smith 1984).

While these psychological models seem to provide better accounts of empirical evidence, they are difficult to implement, and so a majority of AI or Machine Learning implementations are based on the classical model. These implementation are largely designed to exploit the formal syntactic properties of the classical model, and are not concerned with semantic representation issues (see Shavlik and Diettrich 1990, for examples of implementations). They assume a top down decomposition of concepts and are confined to issues related to symbolic encodings with fixed semantics. The paradigm pervades research work in AI and has shortcomings similar to those encountered in other fields such as planning and knowledge representation where cognitive processes are modelled as purely syntactic phenomena; a combination of highly abstract symbolic representations (i.e., the database or knowledge base) and a set of formal rules constraining possible functions over them. The semantic values are fixed beforehand and really function as little more than labels to be interpreted by the user. Though perfectly adequate as logically constructed inference engines operating over well-formed syntactic structures, such models give poor accounts of cognitive processes. As abstract, symbolic representations, they ignore the role of semantics in influencing, altering and constraining concept formation and use. And as a result of this truncated approach such accounts also ignore the role of context. This being the result of treating acquisition of semantics of features and concepts and their subsequent role in cognitive processing is subordinate to the formal (as defined in term of classical logic) syntactic aspect of cognition. Our approach, is aimed at breaking this chain of recursive syntactic definitions, and to do away with the homunculus metaphor which dominates current AI. We will return to a detailed consideration of this issue in the next section.

To a much lesser degree, probabilistic models have also been implemented (Hayes-Roth and Hayes-Roth 1977; Anderson *et al* 1977). Whatever, their individual merits, all these models provide limited accounts of concept formation. The assumption that concepts are defined by sets of features ignores a lot of other factors that also determine the nature of concepts. The manner in which (primitive) features are formed or acquired may significantly determine their role in defining concepts at different levels. Relationship between concepts is probably determined by more than matches between features. The role of contextual factors may alter the relative saliency of features, which could explain how meaning (intension) of concepts vary across situations (Barsolou 1982) or how relationship between concepts is sensitive to context (Tversky 1977).

In the next section, we argue that some of the inadequacies of contemporary models of concepts are due to simplistic and inadequate accounts of *concept formation*. For the sake of clarity the rest of the discussion is confined to the first, basic or primitive stage or level of concept formation.

3. Evaluation of Models of Concept Formation

Why is it useful for an agent or organism to form concepts? Concepts are derived from individual instances but are not confined in their application to just those instances; they can be applied to classify novel objects and events. The process of concept formation is the process of learning about and encoding invariant properties shared by a set of objects or events in the world. An agent with an ability to form concepts would therefore gain in efficiency in interacting with its environment. This is part of the general advantage of phenotype learning over genotype learning. Concepts enhance the efficiency of cognitive processes related to perception, language understanding, action, planning, etc. Without such internal encodings with semantic values (that is, representations of properties of the world) an agent's repertoire of sophisticated sensory-motor and cognitive behaviour would be limited; it would be simply reflexive, based on hard wired reactions to a finite set of stimuli (in the world). The combination of representations and memory (storage system) increases flexibility in behaviour which enhances performance and the ability to adapt to an environment. Accordingly, concept formation and use is part of an agents survival strategy: A proposition that has very important consequences on how meaning is assigned to concepts. Thus biological function plays an important role in our analysis of cognitive processes -a perspective that mainstream psychology has ignored.

So, concepts are very useful. The question is how are they formed? The information processing approach attempts to answer only part of the question: How are similarities and differences between objects and events in the world identified, abstracted and encoded for classification of subsequent instances of similar objects and events? For example, how do you acquire the concept for birds and use it to recognise birds that you have never seen before? We feel a better question would be, what significant processes determine the rendition of continuous, dynamic and random data into discrete, static and deterministic encodings with the appropriate semantics as emergent properties of the agent-environment interaction? Our interest in concept formation is largely focused on the process of constructing (internal) representations, a proper understanding of which would help us overcome the limitations of models that give incomplete accounts of concept formation. From the information processing point of view concepts are assigned meanings which are assumed to correspond to information in the world. That is, the state of affairs in the world determines the nature of representations (the syntactic structure and semantics). Most psychological and AI accounts having made this assumption proceed to model constraints on concept formation to maintain consistency with predefined, fixed feature semantics. Subsequent problems and limitations of models based on this premise have already been outlined in the previous section. Here we will evaluate the efficacy of the basic assumption of the information processing perspective, i.e., information is a property of the world which via transducers determine the formal structure and semantics of internal encodings.

If it is a matter of seeking out and gathering information from the world then it will *not* be easy to motivate the formation of an initial, basic set of features, for the following reasons. Assume that an agent encodes some invariant property in the world. By what process does the agent assign meaning to that encoding? Innate meaning for such primitive feature would be one way of doing so but that is not particularly insightful, and identifying these innate representations is fraught with problems. The feature can be assigned an arbitrary (but consistent) semantic value, but this raises other problems. For instance, arbitrary assignment of semantics renders the idea of information as a property of the world superfluous. Information is supposed to satisfy the requirements of correspondence between the world and internal representations. Arbitrary assignments of semantic values, even when consistent, could not guarantee an isomorphism between information in the world and information in the representations. This problem is confounded by the fact that there is no independent way of verifying the assumed correspondence between internal representation and objects and events in the world. So encoding may have (internally) consistent semantics but there is no way of checking their validity (something similar was pointed out by Wittgenstein 1922).

This problem resolves itself if it is assumed that even if the world contains information, it is not much use as far are its representation is concerned, and arbitrary assignment of semantic values is fine as long as it is consistent. However, even this weaker position assumes that the world is already divided up into discrete objects and events and the task of the agent is to engage in seeking out pattern of regularities. This seems like a reasonable assumption but the perceived regularities could equally be a consequence of perceptual and cognitive processes designed to search for maximal difference between perceived boundaries and minimal difference within them. If so, the assumed correspondence is suspect; if the agent can determine how the world is perceived then its representation may bear little resemblance to how the world actually is. The point is that any account that assumes that the world is something distinct from the agent lays itself open to the charge of being unable to motivate its semantics. One solution would be to treat the interaction between an agent and its environment as a *complete reality* (Maturana 1981); the representation of the world would then be determined by what action an agent is engaged in in a particular environment.

There is a further problem that the information processing view cannot resolve without severely limiting the explanatory power of computational models based on it. In order to construct concepts to represent invariant properties shared by similar objects in the world one needs to know about the correspondence between concepts and the properties they are to represent. But that knowledge of correspondence is only available *after* the relevant concept has been formed. In other words, a concept cannot be formed unless it already exists. Without motivating its semantics independent of the encodings of invariant properties it would be difficult to have concepts corresponding to objects and events in the world. Typically, computational models *avoid* this problem by fixing the number and possible semantics values of features, and putting constraints on their role in concept formation. However, this does not resolve the basic issue of giving an account for emergent phenomena, that is, these models do not account for the occurrence of totally novel feature (or concepts) -otherwise known as the New Term Problem. This holds for all models with deterministic processes and representations defined in terms of rigid, rule based syntactic model, lacking flexibility in assigning semantic values to representations. They cannot account for the affect of context on the use of concepts, and the possible relationship between concepts is highly constrained (usually by some variation of as hierarchical organisation) -a general problem of all top-down purely syntactic computational models of human cognitive processes. Our approach would seek to define concept formation as an emergent phenomenon, the semantic of which are assigned by its function in the interaction between the agent and its environment, that is, a complex adaptive system with emergent properties.

Psychological models provide a better account of emergent phenomena as an outcome of agent-environment interaction (usually couched in the terms of teacher-pupil interaction). However, a detailed account of the cognitive processes involved is not possible without a better understanding of the interaction between behaviour and cognition. Some have looked at possible constraints on how concepts are formed but the usual assumption that concepts have fixed semantics and operate in a strictly deterministic, non-dynamic systems still leaves them without a satisfactory account of how semantic values are assigned to features and concepts. This is the most important limitation of most models of concepts. It is worth stressing that the issue it not that there exists a correspondence between objects in the world and an internal *encoding* of it. That has to be logically necessary in any model of cognition which admits the possibility of internal representations. However, in the information processing approach it is unclear how the encoding is assigned meaning so that it can be said to *represent* some property of the world. The correspondence between objects and encodings though necessary, is not sufficient to explain this process by which encodings are rendered into representations.

Below we describe our approach which is based on a different set of assumptions that motivate a better account of the process of meaning assignment. It does not appeal either to an anti-representational position (see for example Brooks 1986, 1987), or to truth-value semantics (Dowty Wall and Peters 1981). The former claims that a sophisticated sensory-motor control systems interacting with the environment on the basis of data exchange would be sufficient to elicit behaviour. The latter, relies on the idea that if a proposition exits and it is true, and, believed to be true, then that's what determines its meaning. This is fine, except that the determination of truth value of a proposition is often dependent on contextual factors which are not so easy to identify. Harnad's (1990) solution of grounding symbols is aimed at overcoming just these sort of problems. We incorporate the notion of grounding semantics for encodings but also propose a radical shift in focus; abandon the information processing paradigm, and concentrate on investigating cognitive processes as part of the complete interaction between an agent and its environment.

From the above it is apparent that these problems are partly due to a particular research perspective and the limitation of the methodology employed to investigate the process of concept formation. A combination which ignores the variability or flexibility of the semantic values of concepts and their sensitivity to context. In the next section we describe how these problems are resolved when the focus shifts to concept formation as an emergent property of agent-environment interaction.

4. Ecological Approach: Concepts as Emergent Phenomena

To summarise the discussion so far, the problem with concepts is that while they represent abstract properties of instances of object and events they are also flexible and often unpredictable in their application and usage (*cf.* Lakoff 1987). This is because the semantics of concepts is partly determined by context. This problem is analogous to the one that besets purely syntactic, context-free models of natural language processing. Invariably, they fail to account for a subset of data unless a degree of context-sensitivity is admitted. Hanard (1990) has argued that this problem would be overcome if the semantics of symbols are grounded; that is, symbols or categories or concepts are assigned their semantic values during the learning stage which is a consequence of interaction between agents and environment. Basically, this means that semantics should have a central role in any account of natural language processing, and the best way to ensure this is to provide an account of the acquisition of semantics as an integral part of the process of language acquisition. An approach that a number of psychological studies of language acquisition and mother-child interaction tend to support (Narasimhan, 1991). We are in agreement with this approach but feel that it does not go far enough since it still assumes that symbols end up with more or less fixed semantics after the learning phase is over. From the complex adaptive systems' approach this is neither feasible nor desirable for models of dynamic cognitive processes such as concept formation.

In light of the problems associated with an information processing account of concept formation, a totally different approach will be described in this section. It is based on four basic assumptions:

- Abandon the notion that the world contains information (logical or necessary truths) which agents strive to learn and encode as internal representations.

- Dynamic cognitive processes such as concept formation are modelled as part of complete reality. These processes are not seen as formal syntactic phenomena confined to an abstract symbolic level detached from the rest of agents' sensory and affective systems.

- Assume a correlation between invariant properties in the world and internal encodings but do not motivate the semantics on this basis.

- Motivate the representational content of encodings on goal-related task or strategy -encodings have meaning because their correspondence to certain actions and its outcomes has some purpose for agents. This is the emergent process that we are interested in.

A study motivated by these four requirements is analogous to observational studies typically carried out in ethology and certain branches of developmental psychology, except that studying the behaviour of simulations of interactions between agents and their environment would provide a better insight into the nature of these processes because it provides more control over variables that influence learning. Hence it is worthwhile to concentrate on the dynamic behaviour of complete reality in order to investigate complex adaptive systems like concept formation. A more detailed description and justification of this approach can be found in Sejnowski, Koch and Churchland (1988).

Let us consider some of consequences of the assumptions which serve to distinguish our approach from a majority of mainstream Psychological and AI models. First, the world contains no *a priori* information, at least from the agent's point of view -even if it did it is not particularly accessible as it was pointed out in the previous section. Instead we assume that world contains *data*, which the agent renders into information. Hence information can be regarded as a synthesis of agent-environment interaction. There are no *a priori* truths; information about the world is relative and subjective. This perspective is designed to address the more realistic issue of why agents need to gather information in the first place. The information processing approach assumes that information is useful for agents but rarely elaborates how it is useful. In our case we assume that information is useful only so far as it enables the agent to carry out a goal oriented task more efficiently; that is, it saves energy (reduces costs) and increases the overall survival fitness (increases benefit). The information content of the data from the world is determined by the agent according to its past relevant experiences, present perception of its environment, and future goals. The agent is not regarded as perceiving or acting in the *world* but interacting within *its environmental niche.* The environment is thus defined as that which is important to an agent at any given time (so it can vary according to an agents perception, potential for action from a specific spatio-temporal perspective and goals). This is an ecological approach (in the tradition of, among others, Gibson 1982 and Lorenz 1977) focusing on investigating the emergent properties of interactions between agents and environments.

The second assumption is a direct consequence of the first one. If information is actively (subjectively) constructed from data from the world then the agent's cognitive processes, no matter how high level or abstract or symbolic, must be affected to some degree by the manner in which the data is interpreted. Further, no matter how abstract or syntactic an account of cognitive processes can be given, from an ecological point of view such processes are never completely detached from the continuous agent-environment interaction. This point of view diverges from that of grounded semantics, because there is no assumption that the learning process ever comes to an end. A real continuous world, or its subjective perception, an environment, does not stop having any influence on the agent (except when it is dead); the agent is in constant interaction with its environment and actively engaged in constructing discrete internal representations of aspects of environment appropriate in particular context or for specific tasks, that is, making sense of the world according to its needs. And this activity affects not only meaning, or representational contents of encodings (by being updated, modified, or even discarded for better conceptualisations) but also processes which function over representation. Processes like semantic values are not assumed to be fixed (though they may be stable over a long period of time). These changes can be analysed to determine the influence of specific factors but for the present our description is confined to explanation of the general nature of the relationship between action, sensory-motor perception and cognitive processes.

Interaction between an agent and the environment results in information being created, and, this process is continuous. The aim for the agent is to gain a better understanding, or grasp of its environment in order to increase efficiency and viability. This is done by rendering, a continuous, fluid (dynamic) world into a reasonably stable representation of an environment. So the agent is in constant search of invariant patterns or properties in the world. This process never ceases because the world is in a constant state of flux as is the agent's motivation in interacting with it. Hence, the reason for the third assumption, that there is a correspondence between invariant properties in the world and internal encodings of them. Of course, not all invariant properties are encoded; the agent's perception of its environment would select the appropriate pattern (may be by something as simple as trial and error process). Note this does not motivate the semantics; correspondence in itself is not enough to assign semantic values to encodings of invariance.

Finally, consider the forth assumption. The assignment of semantics which has an element of intensionality can be best described with an example. Assume a very simple organism whose survival depends on being able to distinguish between food and non-food in the world. In order to be able to do so it has to learn to make this distinction. Initially this process is confined to trial and error. Eventually, the organism is likely to succeed in finding food. Assuming that this is accompanied with some sort of encoding about the properties in the environment then the organism can assign a simple meaning to it. It does not have to be very complex, as long as it assigns the value that on encountering similar properties in the environment treat it as edible. This would also include the whole repertoire of associated sensory-motor action and goals. This ensures that encodings have semantic values determined by some goal oriented behaviour of the organisms. Hence, the emphasis is on modelling cognition as biological functions (Valera 1986).

Our approach is aimed at providing a detailed analysis of this process, and we intend to do so by investigating simulations of interactions between agents and environment. Perception of data is determined by appropriate visual-spatial sensors (needless to say, there are other, e.g., auditory or olfactory, that also play an important role in concept formation). Since it is assumed that concepts play an important role in the functioning of an agent, any account of them must also give an account of the relationship between concepts and sensory-motor affectors. Hence, we expect to be able to provide some insight into the properties of the perception of the world and action within it: The overall research objective is to explain the nature of simple concept formation as mediated by agent-environment interaction. To do so adequately we believe one needs to be able to explain the detailed process of transition from continuous (noisy), dynamic, variable data to discrete, static, internal states (that is, representation of invariant features) or symbols.

The novelty of our approach lies in adopting a very different research perspective and our intention to utilise more appropriate methodological tools to study and understand the phenomenon of concept formation. The methodology is borrowed from related fields of ethology, and more recently, (computational) neuroethology and Artificial Life. It focuses on two aspects which differentiates it from the usual methodology employed to study the phenomenon of concept formation. First, it is a bottom-up approach, and second, it takes into account the possible effects of sensory-motor behaviour within an environment. Concept formation is not regarded as abstract, formal cognitive process detached from the rest of the organism. Ecological, bottom-up approach coupled with powerful search and learning algorithms provides the opportunity to explore this as a possible determiner of semantics of basic concept formation. This increases the complexity of the phenomenon of concept formation, but even simpler models based on simulated or simple artificial agent can still prove to be extremely insightful as recent work in this field shows (see review by Meyer and Guillot 1990).

We do not subscribe to any particular theory of how concepts are formed, and are more interested in exploring one possible way in which they could be formed. It is assumed that agents construct and instantiate cognitive processes in accordance with their ecological environments. One consequence of our approach is that individual concepts are not independent of each other; the meaning of a concept is necessarily dependent on the meaning of other concepts (Rorty 1989). Both these motivations are a direct result of the limitations of the information processing based accounts described in sections 2 and 3 above. Concepts are not well-defined entities in themselves. Not only are they not independent from one another, they are also formed on the basis of less complex observations, experiences, or even simple associations. Meaning of representations is not independent of it effect and affect on behaviour. The relationship between sensory-motor data from the environment and internal representations of concepts needs to be taken into account because it serves to determine (ground) their semantic values and helps define their role in cognition in terms of other concepts. This can be achieved by a systematic investigation of the process of simple concept formation from a bottom-up perspective. In this approach the focus is on the processes which account for successful searches for invariance in the data about and from the environment. A good grasp of such processes would enable us to develop a model of concept formation based on emergent properties of internal states which can be reliably correlated with interpretable sensory-motor behaviour. It is important to note that this approach is designed to observe learning and behaviour as emergent properties of dynamic processes.

More generally, it is not in the nature of this methodological approach to assume that learning comes to an end at a particular stage, though the incremental learning may diminish to a vanishingly small point. So far such studies have been confined to ethology and developmental psychology. The former concentrates on non-linguistic behaviour and the latter, on mainly language acquisition and development of self-awareness. Note, however that neither approach postulate detailed models of the process of rendering continuous data into discrete internal representations. We aim to do so by developing detailed simulation model and studying actual behaviour of simple (simulated or artificial) autonomous agents. Hence, in this approach the definition of concept formation is slightly different: a more comprehensive one, unlike in models evaluated above.

5. Complex Adaptive Models of Concepts

The overall object of the exercise is to study the behaviour of an autonomous system (agent) in its unstructured environment. The interaction between an agent and its environment can either be simulated or represented by a real robot. In our case we expect to investigate the behaviour in both types of implementation. It has been argued (Hanard 1990) that there are non-trivial differences between these two approaches. Basically, the difference is due to different sources of raw data. In a simulated environment, no matter how realistic, the data input or as *perceived* by the agent is not *direct*, that is, it is not completely continuous or variable. In such a case the degrees of freedom are constrained unlike the case in which the transducers perceive the environment directly. There is, for example, a sense of immediacy which determines the crucial difference between a simulated cockpit and flying a real areoplane.

However, for practical research purposes simulations are not a complete waste of time. They represent a very efficient tool to study the emergence of activity of autonomous systems (agents) in a realistic time scale. What seems to be important at this point, is the notion of autonomy. We use this term in a very strict sense: autonomy is not defined in terms of self-containment, on-board computing facilities and battery supply. An autonomous system determines its activity purely in terms of an internal reference scheme, or internal needs. These global needs are basically pre-programmed but do not therefore determine the exact nature of the behaviour of the autonomous system. In other words, apart from the end goal(s) the system is not programmed to carry out specific tasks. Alternatively it has a general level of flexibility behaviours and adaptive learning to enable it deal with the ecological demands in a dynamic environment.

Reinforcement learning procedure can be used to incorporate these reference schemes into our system architecture. Reinforcement learning takes place as an outcome of the agent-environment interaction. The agent has no knowledge about its surrounding environment, that is, about objects, properties of objects, rules, plans, etc. The learning system interacts with its environment via detectors (sensors) and effectors. Feedback which indicates usefulness or uselessness of an agents behaviour however, is not part of this interaction, as illustrated in the diagram in Figure 1 below.

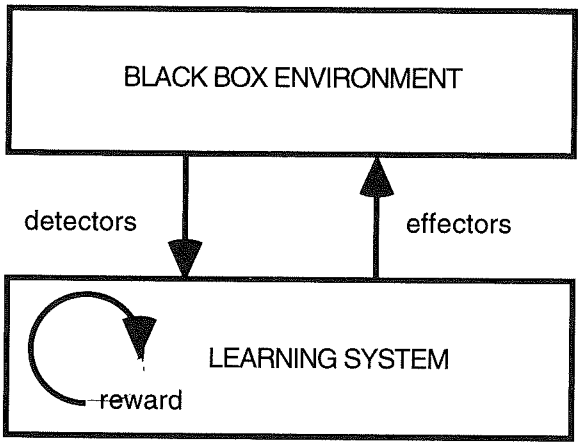

Figure 1: Reinforcement Learning

The detectors receive data from the environment which corresponds to some state of the environment. The effectors act on the state in the environment[4]. The outcome of an agent's action is not known to the agent beforehand. Nor does the environment provide any information on the type of action an agent should carry out; this is determined solely by the reward function. The agent's overall goal is always to maximise reward for any action in all environmental situations. The learning procedures (which is described in greater detail below) enable the agent to build up internal representations of objects, events and plans. This structuring of the environmental data on the basis of internal feedback comprises a dynamic construction process necessary for basic cognitive tasks such as pattern recognition, generalisation, abstraction and concept formation.

5.1. Inductive Learning and Concept Formation

Holland *et al* (1986) state that, at the lowest level of behavioural organisation it is reasonable to assume that "[animals] are capable of applying inferential rules in the relationships between events and that this practice results in the formation of a family of hypotheses to be tested and modified by subsequent experience". The study of induction, then, is the study of how knowledge is modified through use. It is an inferential process that expands knowledge in the face of uncertainty. In this process, probabilistic and statistical inferences are highly relevant. Traditionally, inferential processes are characterised as a process of search through a state space that has an initial state, one or more goals, and a set of operators that can transform one state into another. In (classical) AI, these techniques apply operators to fixed problem representations mainly in the symbolic domain yielding appropriate syntactic structures. Usually, no learning is involved, and the system needs knowledge beforehand to solve a problem (which itself has to be well-defined in terms of initial state, goal-state, and operators). However, this is not a very realistic approach since the class of sufficiently well-defined problem is very small; usually problems are ill-defined, and even worse, arbitrarily ill-defined.

Here we outline our proposal for a naturalistic cognitive system which can process environment input, store knowledge and thus benefit from experience such that it has some basis for action even in unfamiliar situations. While being highly critical of the linguistic-oriented syntactic approaches to reasoning in classical AI, Holland *et al* (1986) focus on pragmatics. They favour representational changes through recategorisation of problem components and by retrieval of associations and analogies. In contrast to classical AI this approach is based on the assumption that people rely on pragmatic reasoning schemes (i.e., clusters of inferential rules) that characterise relations over general classes of objects, event relationships, and problem goals. Thus the focus is on the role of induction in problem-solving behaviour, and, the examination of the relation between goal-directed problem-solving behaviour and induction of new rules. Accordingly, induction is, directed by problem-solving activity, and, based on feedback regarding the success or failure of predictions generated by the system.

Nevertheless, in our opinion this characterisation of problem solving behaviour is still confined to a relatively abstract symbolic cognitive level. It treats cognitive processes as some sort of functions or rules that operate over representations (of knowledge, for example), and thus firmly rooted in information-processing paradigm dominant in AI. In contrast, we apply these inductive processes to behavioural organisation and non-symbolic (or sub-symbolic) representation avoiding any notion of a syntactic level. The focus is firmly on semantics rather than syntactic or pragmatic aspects (though, of course, we do not deny their function in cognitive processes). However, in order to develop autonomous agents capable of flexible and adaptive behaviour the focus has to be on inductive processes based on experiences (via interaction with the environment) that have internal semantics rather than abstraction such as syntactic structures. So emergent behaviour and concepts are matched with internal states with semantic values determined by past experience of the agent. These internal states will have captured regularities in the environment which effectively is their (relatively stable) semantic values. Once learnt, the internal semantics are expected to play a greater role in constructing information from subsequent incoming data.

To summarise, for inductive problem solving to work a cognitive system has to construct a model of the problem space. Effectively the inductive process generates possible solutions (or sub-solutions) and which are revised in light of feedback (based on the outcome of the proposed solutions). Flexibility in the generation of possible solution reflects the level and extent of existing knowledge structures. What kind of computational device could support this type of process of inductive problem solving? Our proposed implementation of Learning Classifier systems (Smith and Goldberg 1990; Holland *et al*, 1986) seem to have the necessary power and flexibility as will be shown in the next section.

5.2. Computational Model and Learning Classifier Systems

A computational structure which supports learning and cognitive capabilities would need to have a the following properties:

- It is modular and parallel in order to represent the many concurrent processes operating on the various signals from effector and detectors

- Individual modules have learning capabilities in order to adapt their behaviour according to outcome of previous action (experience).

- The interaction between the various modules has to be dynamic enough to model the flexibility and plasticity of natural systems.

Such a cognitive architecture (illustrated in Figure 2) serves the main purpose, of controlling the activity of the autonomous agent. Its task would be to learn simple behaviour in response to certain stimuli in the environment. The agent will be pre-disposed to elicit some sorts of behaviour and avoid others. This aspect will be incorporated by introducing a set of desirable goals that can be achieved by appropriate behaviour. Note that such goals are only implicitly stated by means of an evaluation function which determines the relative usefulness of an action (on the bases of its outcome) and rewarding (or punishing) it appropriately.

This ensures that only appropriate (that is, internal states which enable the agent to perform appropriate behaviour) kinds of concepts are learnt. With this combination of induction and adaptive reinforcement learning an agent is expect to construct internal world models as a result of its interaction with the real world. Such a model may include knowledge about its spatial position and the relative positions of its sensory organs and motor controls, though the exact nature of it will only become clear as a result of our proposed research. For a more detailed description of the proposed computational architecture see Schnepf (1991) and Patel and Schnepf (1991).

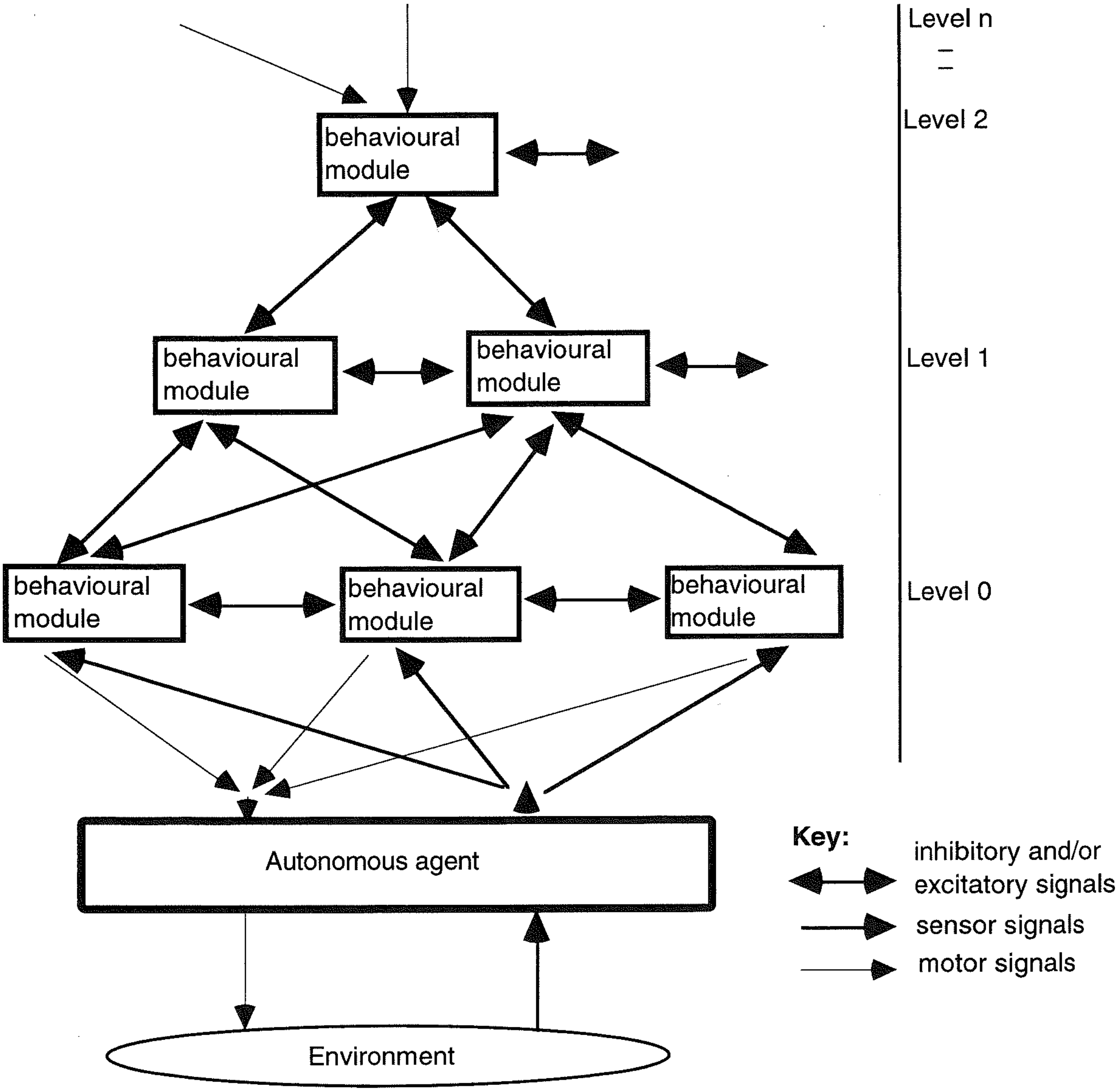

Figure 2: Behavioural organisation in an artificial agent

Learning classifier systems (LCS's) provide the adaptivity necessary for reinforcement learning and inductive processes. Here we give a brief account of

LCS's and how they can be utilised to model knowledge abstraction, generalisation, and concept formations. (See Golberg, 1989 for further details). A classifier system is a kind of rule-based or production system. There are several reasons for choosing rules as the most basic representation for inductive processes. First, it is an efficient way of specifying system behaviour. Second, their modularity means that new rules do not change existing rules. Third, rules can describe transitions from one system state to another. Further for theoretical purposes a set of rules can be regarded as mental model of experience.

The rules, or *classifiers*, consist of a condition and action part which are formed by a very low level syntax usually consisting of a three-letter alphabet, *0,1* and *# (don't care).* Messages (inputs) from the environment are matched against the condition parts of classifiers. Classifier systems process rules in parallel -an important feature of such systems. If more than one classifier matches the environmental message, an additional parameter, that is, each matching classifier's *strength*, is used to determine which classifier is applied or acted upon. A feedback mechanism (internal or external) evaluates the efficacy of the action by comparing the environmental and the agent's internal state before and after the action, and then allocates positive or negative feedback (reward which is a measure of their strength) according to whether it was a beneficial (in terms of the desired goal) action or not.

To generate new rules an evolutionary algorithm, typically a *genetic algorithm,* is applied at regular periods during learning. New rules are generated by recombining existing rules to mimic process of natural evolution -so rules with high strength have proportionally higher probabilities of being selected for generating new rules. This is a powerful mechanism for enhancing the flexibility and adaptivity of the agent, and thus overcoming some of the limitations of highly constrained rule production system. From a different perspective the operation of the evolutionary algorithm ensures a more realistic and exhaustive search (mediated by evolutionary selective pressures) for rules appropriate for solving a problem or learning a task via interaction with the environment.

In standard classifier systems, adaptive rule learning is performed by one processing unit, i.e., the classifier system itself, having clusters of rules as the largest group of situation-action pairs related to each other. Consequently, the complexity of action sequences (we refer to this as behavioural sequences) to be learnt and performed afterwards is not very high. Additional disadvantages of such systems arise from the apportionment of credit problem over long sequences of chained rules and the problem of concept formation (Wilson and Goldberg 1989; Booker 1989). Therefore we propose to implement a cluster of classifier system operating in parallel to sustain more complex behavioural sequences to support concept formation processes. The task is effectively modularised such that each LCS learns a smaller (and perhaps simpler) set of rules. Such LCS's modules (labelled, behavioural modules in figure 2) are organised by the overall learning system which adds an extra dimension to the organisation of rules and rule clusters as described above. A behavioural module reflects to some extent the structure and functionality of a learning classifier system as it incorporates the three main components of such a system: rule and message system, apportionment of credit system, and evolutionary algorithm. The basic task is identical to what a classifier system performs: Environmental stimuli are presented to each behavioural module, matching classifiers are competing, finally one rule becomes selected and fires.

This sort of computational model can be instantiated on a robot to model concept formation via agent-environment interaction. The basic idea is to get the agent to become sensitive to certain regularities in the incoming data. The general properties of the computational model enables an agent to form simple sensory invariants and to relate these invariants to others (in higher level modules) to configure more abstract representation, that is concepts. What sorts of regularities are learnt (and concepts formed) is expected to be determined by the feedback provided by the evaluation mechanism. So appropriate behaviour will depend on consistent internal representations corresponding to regular (invariant) properties of environment and sensory-motor capabilities. By observing the agents behaviour in terms of its ability to achieve the goals and identifying corresponding internal states we expect to be able to get a better idea of how concepts (about the world) are formed.

6. Conclusion

The set of tools and our methodological approach will enable us to observe the relationship between sensory-motor perception and action within a specific environment. The relationship between sensors and effectors, as mediated by internally represented concepts will be explored. We hope to describe the nature of properties and processes involved in invariant features extraction from raw continuous data, and the manner in which this information affects sensory-motor behaviour in increasingly complex scenarios. Our novel but simple approach is not expected to give an account of the effect of memory or past experience on behaviour in a particular context. Apart from making the issue too complicated we feel that the primary goal of understanding the nature of concept formation as an emergent property of a complex adaptive system needs to be well accounted for before attempting to explain n even more complicated system.

* Progetto di Intelligenza Artificiale e Robotica
Dipartimento di Elettronica e Informazione
Politecnico di Milano

Notes

* ERCIM Research Fellow funded by EC Human Capital Mobility Programme.

[1] An *Artificial* agent or organism refers to a computational structure which reflects to some extent the characteristics of real organisms in terms of autonomy, desires, goals, needs etc., and which has the ability to sense and act in its (evntually computational) environment according to these characterisitcs.

[2] We make an important distinction between the world and the environment. The latter is a subjective, agent-oriented persective on the former; an agents environment is one possible subset of the world. This distinction emphasises the importance of context in cognitive processes, and that role of context is determined not by absolute properties of the world but by subjective perception of the agent.

[3] Of course, this is not the case if the model is based on the assumption of *a priori* semantics for basic features. For obvious reasons such models are not of any interest here.

[4] The agent is considered to be part of the environment itself, as activity performed by it can change both the environmental state as well as the agent-environment relation.

BIBLIOGRAPHY

Anderson, J.A., Silverstein, J.W., Ritz, S.R. and Jones, R.S. (1977), "Distinctive features, categorical perception and probability learning: Some application of a neural model", *Psychological Review* 84, 413-451.

Barsolou, L.W. (1982), "Context-independent and context-dependent information in concepts", *Memory and Cognition* 10, 82-93.

Booker, L.B. (1989), "Triggered Rule Discovery in Classifier Systems", in *Proceedings of the Third International Conference on Genetic Algorithms*, Fairfax, USA, 258-267.

Brooks, R.A. (1986), "A Robust Layered Control System for a Mobile Robot", *IEEE Journal of Robotics and Automation*, RA-2(1), 14-23.

Brooks, R.A. (1987), "Intelligence without representations", in *Proceedings of Workshop on Foundations of Intelligence*, MIT, Endicott House.

Collin, A. and Loftus, E.F. (1981), "A Spreading Activation Theory of Semantic Processing, *Psychological Review* 82, 407-428.

Dowty, D.R., Wall, R.E. and Peters, S. (1981), *Introduction to Montagues Semantics*, London: Reidel.

Feldman, J.A. (1986), *Neural Representation of Conceptual Knowledge*, University of Rochester, Dept. of Computer Science, TR189.

Gibson, J.J. (1982), *Wahrnehmung and Umwelt*, Urban & Schwarzenberg.

Goldberg, D.E. (1989), *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.

Harnad, S. (1990), "The Symbol Grounding Problem", *Physica D* 42, 335-346.

Hayes-Roth, B. and Hayes-Roth, F. (1977), "Concept learning and the recognition and classification of exemplars", *Journal of Verbal Learning and Verbal Behavior* 16, 321-338.

Hinton, G.E. (1986), "Learning Distributed Representations of Concepts", in *Proceeding of Eighth Annual Conference of the Cognitive Science Society*, Amherst, MA.

Hinton, G.E. (1990), "Mapping Part-Whole Hierarchies into Connectionist Networks", *Artificial Intelligence* 46, 47-75.

Holland, J.H., Keith, J.H., Richard, E.N. and Paul, R.T. (1986), *Induction*, Cambridge, MIT Press.

Holyoak, K.J. and Glass, A.L. (1975), "The role of contradictions and counterexamples in rejection of false sentences", *Journal of Verbal Learning and Verbal Behavior* 14, 215-239.

Keil, F. (1989), *Concepts, Kinds and Cognitive Development*, Cambridge, MIT Press.

Lakoff, G. (1987), *Women, fire and dangerous things: What categories reveal about the mind*, Chicago, University of Chicago Press.

Lorenz, K. (1977), *Die Ruckseite des Spiegels*, dtv.

Maturana, H. (1981), "Autopoiesis", in M. Zeleney (ed.), *Autopoeisis: A Theory of Living Organisation*, New York, North-Holland.

Medin, D.L. and Smith, E.E. (1984), "Concepts and concept formation", *Annual Review of Psychology* 35, 113-138.

Meyer, J.-A. and Guillot, A. (1990), *From Animals to Animats: Everything you wanted to know about the simulation of adaptive behaviours*, Paris, Ecole Normale Superieure, Technical Report BioInfo-90-1.

Narasimhan, R. (1991), "The Ethological Approach to the Study of First Language Acquisition by Children", in W.A. Ainsworth (ed.), *Advances in Speech, Hearing and Language Processing*, Vol. II. London, Jai Press Ltd.

Patel, M.J. and Schnepf, U. (1991), "Concept Formation as Emergent Phenomena", in *Proceedings of the First European Conference on Artificial Life: ECAL91*, Paris.

Rorty, R. (1989), *Contingency, irony and solidarity*, Cambridge, CUP.

Rosch, E. (1975), "Principles of Categorisation", in E. Rosch and B.B. Lloyd (eds.), *Cognition and Categorisation*, Hillsdale, N.J., Erlbaum.

Rosch, E. and Lloyd, B.B. (1978), *Cognition and Categorisation*, Hillsdale, N.J., Erlbaum.

Rosch, E. and Mervis, C.B. (1975), "Family resemblances: Studies in the internal structure of categories", *Cognitive Psychology* 7, 573-605.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M. and Boyes-Braem, P. (1976), "Basic objects in natural categories", *Cognitive Psychology* 8, 382-439.

Schnepf, U. (1991), "Robot Ethology: a Proposal for the Research in Intelligent Autonomous Systems", in *Proceedings of the International Conference on the Simulation of Adaptive Behavior: SAB90*, MIT Press/Bradford Books.

Sejnowski, T.J., Koch, C. and Churchland, P.S. (1988), "Computational Neuroscience", *Science* 241, 1299-1306.

Shavlik, J.W. and Dietterich, (eds.) (1990), *Readings in Machine Learning*, San Mateo, Morgan Kaufmann.

Smith, R.E. and Goldberg, D.E. (1990), *Reinforcement Learning With Classifier Systems*, IEEE.

Smith, E.E. and Medin, D.L. (1981), *Categories and Concepts*, London, Harvard University Press.

Tversky, A. (1977), "Features of similarities", *Psychological Review* 84, 327-352.

Varela, F.J. (1986), *The Science and Technology of Cognition: Emerging Trends*, mss.

Wilson, S.W. and Goldberg, D.E. (1989), "A Critical Review of Classifier Systems", in *Proceedings of the Third International Conference on Genetic Algorithms*, Fairfax, USA.

Wittgenstein, L. (1922), *Tractatus Logico-philosophicus*, London, RKP (1961 Edition).

Wrobel, S. (1990), *Concept Formation in Man and Machine: Fundamental Issues*, mss, Bonn, GMD.