Time for AI and Society

PROCEEDINGS OF THE AISB'00 SYMPOSIUM ON STARTING FROM SOCIETY - THE APPLICATION OF SOCIAL ANALOGIES TO COMPUTATIONAL SYSTEMS

17th-20th April, 2000
University of Birmingham

AISB'00 Convention

17th-20th April 2000

University of Birmingham
England

Proceedings of the AISB'00 Symposium on

Starting from Society - The Application of Social Analogies to Computational Systems

Published by

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour

United Kingdom

http://www.cogs.susx.ac.uk/aisb/

ISBN 1 902956 13 8

Printed at the University of Birmingham, Edgbaston, Birmingham B15 2TT, England.

Contents

The AISB'00 Convention

The millennial nature of current year, and the fact that it is also the University of Birmingham's centennial year, made it timely to have the focus of this year's Convention be the question of interactions between AI and society. These interactions include not just the benefits or drawbacks of AI for society at large, but also the less obvious but increasingly examined ways in which consideration of society can contribute to AI. The latter type of contribution is most obviously on the topic of societies of intelligent artificial (and human) agents. But another aspect is the increasing feeling in many quarters that what has traditionally been regarded as cognition of a single agent is in reality partly a social phenomenon or product.

The seven symposia that largely constitute the Convention represent various ways in which society and AI can contribute to or otherwise affect each other. The topics of the symposia are as follows: Starting from Society: The Application of Social Analogies to Computational Systems; AI Planning and Intelligent Agents; Artificial Intelligence in Bioinformatics; How to Design a Functioning Mind; Creative and Cultural Aspects of AI and Cognitive Science; Artificial Intelligence and Legal Reasoning; and Artificial Intelligence, Ethics and (Quasi-)Human Rights. The Proceedings of each symposium is a separate document, published by AISB. Lists of presenters, together with abstracts, can be found at the convention website, at http://www.cs.bham.ac.uk/~mgl/aisb/.

The symposia are complemented by four plenary invited talks from internationally eminent AI researchers: Alan Bundy ("what is a proof?"- on the sociological aspects of the notion of proof); Geoffrey Hinton ("how to train a community of stochastic generative models"); Marvin Minsky ("an architecture for a society of mind"); and Aaron Sloman ("from intelligent organisms to intelligent social systems: how evolution of meta-management supports social/ cultural advances"). The abstracts for these talks can be found at the convention website.

We would like to thank all who have helped us in the organization, development and conduct of the convention, and especially: various officials at the University of Birmingham, for their efficient help with general conference organization; the Birmingham Convention and Visitor Bureau for their ready help with accommodation arrangements, including their provision of special hotel rates for all University of Birmingham events in the current year; Sammy Snow in the School of Computer Science at the university for her secretarial and event-arranging skills; technical staff in the School for help with various arrangements; several research students for their volunteered assistance; the Centre for Educational Technology and Distance Learning at the university for hosting visits by convention delegates; the symposium authors for contributing papers; the Committee of the AISB for their suggestions and guidance; Geraint Wiggins for advice based on and material relating to AISB'99; the invited speakers for the donation of their time and effort; the symposium chairs and programme committees for their hard work and inspirational ideas; the Institue for Electrical Engineers for their sponsorship; and the Engineering and Physical Sciences Research Council for a valuable grant.

John Barnden & Mark Lee

One[1] of the cognitive processes responsible for social propagation is *social learning*, broadly meant as the process by means of which agents' acquisition of new information is caused or favoured by their being exposed to one another in a common environment.

Social learning results from one or another of a number of social phenomena, the most important of which are social facilitation and imitation (see fig. 2[2]).

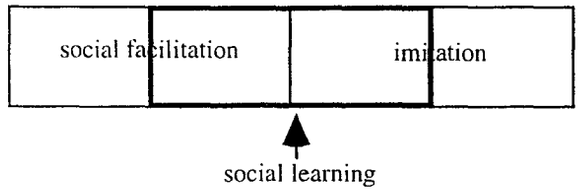

Fig. 2: Social learning

In this paper, a general notion of social learning will be defined and the main processes which are responsible for it, namely social facilitation and imitation will be analysed in terms of the social mental processes they require.

The rest of the paper will be organised as follows. In the next section, where some classical definitions of social learning are analysed, a systematic and consistent treatment of these notions is shown to be missing. In the successive section, a general notion of social learning is introduced and the two main processes which may lead to it, social facilitation and imitation, will be defined as different steps on a continuum of cognitive complexity. In the final section, the utility of a the present approach will be discussed.

The analysis presented in this paper draws upon a cognitive model of social action (cf. Conte & Castelfranchi, 1995; for a synthesis, see Conte, 1999). The agent model which will be referred to throughout the paper is a cognitive model, endowed with mental properties for pursuing goals and intentions, and for knowledge-based action. Therefore, some notions drawn from the formal study of mental states will also be employed.

To be noted, a cognitive agent is not to be necessarily meant as a *natural* system, although many examples examined in the paper are drawn from the real social life of humans. Cognitive agents may also be artificial systems endowed with knowledge and the capacity for reasoning, planning, and decision-making. The interesting question concerning artificial systems is, what are the mechanisms which must be implemented at the agent level to enable them to learn from one another? Are the mechanisms allowing agents to learn from their physical environment sufficient for them to learn also from or perhaps through their social environment? If not, which additional properties are needed? And, earlier than this, what does social learning mean, which social phenomena are referred to by this notion?

2 Classical definitions

It has been observed (Laland & Odling-Smee, 1999) that the term social learning describes a "ragbag" of heterogeneous phenomena, with a variety of functions. A systematic treatment of these notions is still wanted.
In the social psychological literature (Bandura, 1977), social learning is seen as people learning through the observation of attractive and consistent social models. By observing their social models and recording when these apply reinforcing mechanisms, people learn to reinforce themselves (self-reinforcement) to do what others have reinforced, and abstain from doing what others have punished.

This apparently simple and elegant theory has many drawbacks. First, it is exclusively focused on the mechanism of *reinforcement*. What about learning from social models who are unaware about their role and therefore unable to apply prize or penalty?

Second, social learning is essentially meant as a mechanism of *emulation*, which implies the corresponding motivation to look and behave like attractive social models in order to be seen as comparable or similar to them and obtain their approval. What about learning independent of the reputation of the others? Cannot people learn from others without emulating them? Is it possible to formulate a general notion of social learning which includes but is not reduced to emulation?

Finally, given the above notion of social learning, how to distinguish it from imitation? So far, imitation has not been clearly defined. It has been often if not exclusively defined as a behavioural phenomenon. In the typical behaviourist view, recently reworded by Blackmore (1999), imitation is defined as copying a new *form of behaviour*. But what is a new form of behaviour? As a long line of psychological thought has shown (see Plotkin, 1994, for a clear summary), behaviour is essentially a goal-directed or end-directed activity. In this sense, coughing is not behaviour, unless one coughs to signal disappointment or disapproval. When one learns to raise one's arm when meeting another (known) agent, one learns a new behaviour, although the movements involved in such a behaviour were already part of one's action repertoire. In this sense, learning a new *form* of behaviour by imitation means learning a use or meaning (read, goals) which may "in-*form*" (Plotkin, 1994) a given activity. It then becomes apparent that imitation leads to agents' acquiring novel behavioural in-formation from others, and therefore implies their capacity to draw such information from observed behaviours. Furthermore,

[1] Another important phenomenon is social control, by means of which agents influence one another to comply with intra-groups norms.

[2] Neither social facilitation nor imitation do necessarily lead to the agent's acquisition of novel information: the former, as often found in the social psychological literature, may lead to improved performance and goal emulation. The latter, especially when aimed to reuce social distance and increase conformity, may lead to comply with standards and norms already acquired by the agent.

is, "never stop under a tree when it rains". In such a case, one agent learns from another without the latter's behaviour to propagate. The example shows that social facilitation is very close to an even more elementary type of learning: the observer might infer the same lethal effect even by watching the tree being struck by lightning. However, as shown by the ethological literature, learning is enhanced by observing the effects of actions or external events on *conspecifics:* what happens to a conspecific will (be expected to) concern me more than what happens to a tree. However, this point deserves further clarification: it is not only the outcome of the process (struck by lightning) that which is bound to elicit the learning process (don't stop under a tree when it rains), but also the *process leading to interpret* the outcome as relevant for oneself (the entity observed).

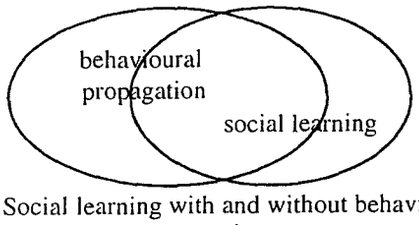

Fig. 3: Social learning with and without behavioural propagation

Social facilitation is a very elementary type of social learning, in which the beneficiary does not necessarily attribute the other any goals or other mental states. In the above example, the same effect might have been probably achieved by the observer, had it seen a piece of wood, rather than a fellow agent, struck by lightning. Of course, if the input comes from an agent, the stimulus to the observer's inference and the chance that she gets new information out of it are higher (since it is higher the probability that the event observed has effects on the observer's goals, which may overlap with the input agent's to some extent).

There are different types of social facilitation phenomena, according to the role played by the input agent (which will be called, the Source) in the Observer's learning process and in its representations. In social facilitation, S may operate as a

a) *pointer or "bookmarker"*: S acts in such a ways as to increase the chances that O perceives a given event, which triggers O's learning process. This is essentially what ethologists call *local enhancement.* As an example, while running after S, O discovers a new region which she[5] had never realised before. Here, there is no need for S to transmit his own beliefs to O (S might simply escape from O in an unknown direction). O, in her turn, does not acquire a new piece of knowledge by reconstructing S's mental states, nor by observing the effects of a given event on S's fate. S is a mere accidental cause of O's discovery. S acts as a sort of bookmarker or pointer. Here, O learns a new section of the world map.

b) *Qualifier*: S's features may characterise a given environment, and help O characterise or identify it. Suppose I am in a foreign country and badly need a restroom, but cannot tell from the written signs which is the ladies' and which is the gentlemen's toilet. One possible solution is wait and see which way will take the next newcomer of either gender. Interestingly, the social cognitive process which occurs is the same but leads to alternative behaviours: if the newcomer belongs to my gender I will act alike; if he is of the opposite gender, I will take the alternative way.

c) *Activator*. This is shown by the example of milk bottle top opening in British tits (Hinde & Fisher, 1951). Let us see how Laland and Odling-Smee (1999: 6) discuss this example: "These birds learned to peck open the foil cap on milk bottles... . Hinde and Fisher found that this behaviour probably spreads by local enhancement, where the tits' attention is drawn to the milk bottles by a conspecific, and after this initial tip off, they subsequently learn on their own how to open the tops". However, Hinde and Fisher's explanation is insufficient. The learning process is facilitated by S in a double way: S draws O's attention on a given object, which possibly "activates" O's goal of manipulating it, and therefore leads O to exhibit the same behaviour as S. Here, propagation occurs. However, there is no need that O actually represents S as a "manipulator", nor, a fortiori, that O attributes S any capacity or mental states. S points to a new object which might activate a built-in routine for manipulation. An analogous example is offered by the acquisition of dietary preferences among rats (Galef, 1996; see again, the discussion of this example in Laland & Odling-Smee's, 1999: 6), which prefer "to eat foods that other rats have eaten".

d) *Belief-holder*: a subset of S's inferable beliefs may help O to identify and understand the environment. An interesting example of this phenomenon is offered by people's recognising a given (social) setting by observing others' behaviour: if someone is standing on the edge of the sidewalk, it is probably there where the bus stops. In such cases, O resorts to her pre-established beliefs about S (pedestrian): people standing up motionless in the street usually are waiting for someone or something. Interestingly, O may have a pre-existing goal (taking the bus), which S helps her to achieve by marking how to verify its preconditions (find the place where the bus stops). Alternatively, this goal may be activated by O's perception of S's behaviour and by inferring the associated mental states (O is walking to destination, but since she understands that a bus-stop is near, she may get on the next bus). In such a case, social facilitation allows for social propagation: a given (set of) belief(s) travels from S to O. Indeed, O decodes S's beliefs from his behaviour and incorporates them into her knowledge base (unless she finds evidence that S is wrong or her inference is incorrect).

[5] In the remaining of the paper, explicit referance will be made to human agents to facilitate the reader's understand the reference of pronouns (O will be a female agent, and S will be a male agent). However, some, if not all the processes analysed may occur among non-human organisms and even among non-natural systems.

e) *Experimental "testbed"*: this is shown by the example of looking for a shelter from rain. By observing what happens to S, O learns to avoid trees. Here, O learns a negative effect of a known plan of action. Examples of this sort abound in social life: agents not only observe and learn given behaviours from one another, but also avoid the costs of a direct experiment, and learn the positive or negative (side-)effects of current plans/procedures etc.

f) *Subject of norms, standards, conventions:* S's behaviour may indicate existing standards, norms, and conventions. Independent of whether O will decide to accept or reject them, and of whether to comply with them or not, others may be a fundamental source of information about formal or informal norms, customs, habits and any other factor of regulation of one's (social) conduct. More basically, think of O as an external observer, an anthropologist or ethnographer. She may learn a lot about how a given society is organised, differentiated, what are its social hierarchies, etc. from the behaviours of the society's members. But even independent of standards and norms, O may learn a lot about social categories, reputation, roles, etc. by observing how agents interact with and react to one another.

To sum up, social facilitation is a mechanism by means of which a given agent updates her knowledge base, including social and pragmatic knowledge, by observing others, their features and behaviours, and possibly (but not necessarily) by inferring their mental states.

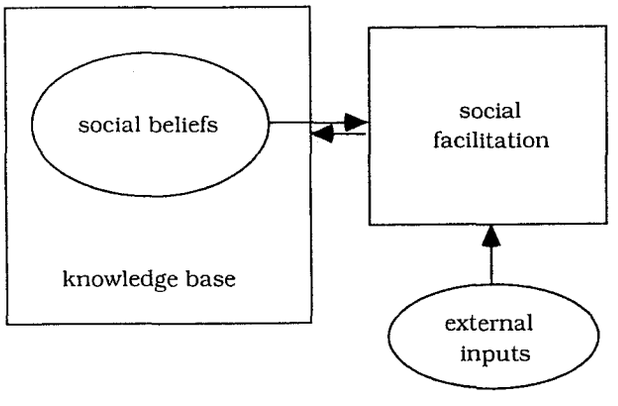

Fig. 4: Inputs and outputs of social facilitation

5.2 Imitation

In the previous section, social facilitation has been defined and described as a type of social learning in which the learning agent (O) updates her knowledge base by perceiving the relationship between another agent (S) and its physical or social environment. Such relationships may (or may not) include the effects of S's behaviour on the environment, and/or the effects of the environment on him (and possibly his achievements). In social facilitation, O receives information relevant for her current or potential goals by observing S in a common environment. Consequently, O forms beliefs/perceptions about S (from which she acquires novel information), but her goals do not mention S. S simply plays the role of an implicit, undeliberate, even accidental indicator or even informant about the environment.

In this section, another step of the social learning process is analysed, namely imitation. Imitation is here defined as a phenomenon of social learning in which the learning agent is ruled by two social *goals* concerning S (a social goal being defined as a goal which mentions another agent's mental states; cf. Conte, 1999):

a) know what S does, how he behaves, how he looks, etc. in order to find out standards, rules, or simply means to achieve her own goals. O's social goal is a means for O to reach another goal of hers. The latter might be specific or generic. For example, O may not know how to use the silverware in a fancy restaurant. She then looks around to see what her fellows do with them.

b) adopt S's goals and/or other mental states and possibly the consequent behaviours, *as long as* O believes that S is an appropriate or adaptive model in a given domain. In the formal treatment of mental states, a goal *relative* (Cohen & Levesque, 1990) to a given belief is a persistent but conditioned goal, that is, a goal which is pursued as long as it is found either unfeasible or already achieved (persistent) or unless the belief associated to it is revised or retreated (conditioned). In the case of imitation, the goal is relative to O's *social* belief: O imitates another agent as long as she believes that it is useful and convenient to do so, namely as long as the other shows an appropriate or adaptive behaviour, looks, style under given circumstances.

> *Imitation is a behaviour ruled by the goal that a given agent (O) be-like or act -like another agent M (which stands for Model), as long as M is (perceived as) a suitable model under a given circumstance.*

The main difference between social facilitation and imitation is that in the former case, O has social beliefs about S, from which O obtains relevant novel information. In imitation, instead, O pursues a number of social goals with regard to M, relative to her belief that M is a good model. These goals actually suggest interesting operational criteria for a model of imitation: if a system is ruled by a goal *relative to* a given belief (Cohen & Levesque, 1990), the system will have to (a) check the truth value of its current belief - in the case of imitation, it will repeatedly monitor (i) M and his doings, (ii) how good (e.g., adaptive or successful) M is as a model; (b) the relativized goal is a persistent but conditioned one - in our case, O will persist in imitation as long as she believes M *is* a good model.

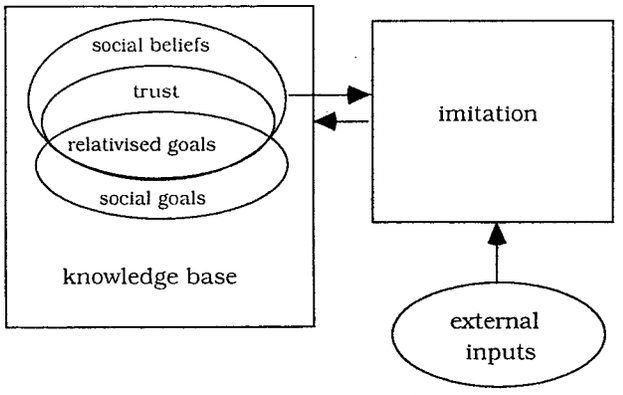

Fig. 5: Mental States in Imitation

5.2.1 Imitation: Goal-Directed or Goal-Oriented Behaviour?

The preceding definition might raise an important question: what type of goals are implied by imitation? What does it mean that O is ruled by a set of social goals? To what extent can this notion of imitation be used, which systems can it be referred to? More explicitly, which mental capacity (or complexity) is required for a system to exhibit and be attributed this type of behaviour? What is the relationship between Mitchell's clause (4) (that behaviour C is *designed to* be similar to M) and the definition here provided?

The question is a difficult one, which will find no conclusive answer here. However, it is at least necessary to recall that goal-directed behaviour (Plotkin, 1994, in line with the Piagetian definition) includes goal-governed behaviour (cf., Conte & Castelfranchi, 1995, ch. 8), that is, behaviour selected by the evolutionary process *to achieve* a given effect, without this effect being necessarily represented as an explicit goal in the imitator's mind. Consequently, imitation among some species might have been selected as one agent's *monitoring of given others* (e.g., parents) in order *to be-like or act-like them* as long as they are (perceived as) *good models* under given circumstances. What is a good model in evolutionary terms? How can "good" models be biologically transmitted without bringing into play a sort of Lamarckian evolution? This is a difficult question which ethological literature can help to answer. For many species, any conspecific has been selected as a good model (read, better than a non-conspecific) at least relative to given types of behaviour (e.g., dietary rules). For other species and in other behavioural contexts, adults have been selected as "good" models (infant chimps imitating adults foraging for termites using stalks), or more specifically "first-perceived" adults (imprinting). To adopt the dietary preferences of conspecifics, in such analysis, is attributable to imitation if a set of operational criteria, corresponding to the relativized goals mentioned above, are met. To define such criteria is beyond the scope of the present model. However, the notion of relativized goal can contribute to formulate them. A behaviour governed by a relativized goal is a behaviour persistent but conditioned to a given belief (that M is a good model): (a) imitating different sets of models (e.g., conspecifics Vs parents) for different domains (dietary preferences vs. problem-solving heuristics) is a necessary but insufficient criterion; in order to check conditioned persistence, one should probably (b) check whether the animal persistently monitors its model (like for example, in imprinting), (c) check whether imitation is conditioned to the model's success (e.g., abandon a given diet learned from a conspecifc after the latter's death).

5.2.2 Why Imitate...

There are several main reasons for O to imitate M,

a) Know what to do: in many (social) circumstances, agents may ignore the efficient or more convenient procedures, rules, plans to achieve their goals. Sometimes, they may even ignore which goals can be achieved, or which activities are feasible or safe, etc.. For example, while taking a walk in an unknown city, I follow the main stream of pedestrians simply because I do not know where to go to, which neighbourhood is safe enough, etc..

b) Comply with social standards and norms. This is very close to social facilitation in the sense that O observes others to infer information about the world from their behaviours. But what is characteristic about imitation is that O is interested not only to find out what the norms are, but also to see how others react to them, to what extent they keep them into account, which ones are applied and which ones are instead ignored. By this means, O learns both relevant *information* about the (social) world, and relevant (social) *conducts*.

c) Fulfil given roles. Sometimes, imitation is prescribed. For example, parents are models that children *must* follow (this has obvious biological forerunners, such as imprinting).

d) Compensate one's inadequacy, ignorance, inexperience in problem solving. If I do not know how to use the chopsticks, I will probably take the customers of a Chinese restaurant, or my friend Yang, as suitable models to observe carefully in order to reproduce their behaviours.

e) Avoid complex problem solving or risks. Sometimes, I am too uncertain about the consequences of a given conduct, and rather than sustaining the cost of a direct experience, I decide to shelter under someone else. For example, if I am alone, I may stop on the edge of a slope, but I may risk to follow someone who decides to proceed.

f) Reduce social distance. At times, people imitate others to conform to them, thereby reducing the social differences within the group. In these cases, O's goal to adopt M's goals, preferences and behaviours is not relativized to her belief that M is an appropriate model, but rather to her belief that M persists in having those goals, preferences, etc.. In its most extreme form, this leads to O following M whatever he does.

g) Share responsibility, commitment, etc. and their costs. In social impact theory, a crucial factor of explanation of bystanders' intervention in emergencies

is precisely the existence of other bystanders. Indeed, the more they are, and the less cohesive the group, the less efficient or timely the intervention of bystanders during emergencies. Why? As Latané and Darley (1970) convincingly explained, people want to share the responsibility of both the decisions involved (whether or not to define the episode as an emergency, and whether or not to intervene) with others. No one wants to find herself in a "vulnerable", isolated, position. Consequently, each one checks what others do and how they react. Of course, this generates a bottleneck: each waits for another to make the first move, which no one will then undertake.

Generally speaking, imitation appears as a short-cut in problem solving and planning. O minimises the costs she should otherwise invest in these activities by accepting others' outputs. Of course, a trade-off may be envisaged here: on one hand, O reduces her own costs, on the other she might increase them by following others' conducts which later might appear useless or risky. Indeed, imitation implies delegation and ultimately trust: O implicitly delegates others to do (part of) the job she should do. She must trust M to some extent. Consequently, imitation leads to another, intrinsically social type of problem-solving and reasoning: whom to trust, and about what? How to tell when someone is trustworthy or reliable, how to tell that his conduct is adaptive and that it is then reasonable to follow it? To put it otherwise, when do agents perceive themselves as adequate in problem solving and when, instead, do they prefer to delegate this to others?

5.2.3 Whom to Imitate

This question is closely related to the role of trust: O imitates M when she trusts M. But to what extent should she trust him, with regard to which competencies or characteristics?

First, imitation may have more or less domain-specific target. This is because trust is relative to specific contexts and domains of competence: I will certainly look at Yang in using chopsticks, but may have no good opinion of his command of the English language and therefore refrain from imitating him in such a context. In addition, imitation may be individualised or not: I may decide to look at my friend Jenny rather than John, because all considered, I trust her competence, problem solving capacities, etc., to a higher extent than I trust his. On the other hand, I may want to look at any colleague who obtained a promotion in the last two years.

Another important dimension, which is related to trust but different from it, is the goal of imitation: a youngster will feel more likely to find models in her own age co-hort than in others. Here, the goal is not problem-solving but reducing one's social distance from a given social aggregate. Therefore, the target of imitation is any agent who is a good representative, a typical exemplar of that aggregate. Obviously, prototypes are *trusted* to possess the characteristics which are essential to the category of reference.

Finally, imitation may be based upon observable frequencies: in many cases, the more frequent a given behaviour and the more it is target of imitation. This has at least three reasons:

a) first, the more frequent a given behaviour, the more it is perceived as rational, in the sense of independent of subjective, idiosyncratic preferences and biases

b) the more frequent a given behaviour and the more it is perceived as one which as proved the fittest (selected by success)

c) the more frequent a given behaviour and the more it is perceived as prescribed, or even mandatory.

5.2.3 ... and What to Imitate

Unlike the classical view of imitation as a strictly behavioural notion, imitation is here seen as a special case of intelligent social behaviour, in which the Observer intends or is designed (to use Mitchell's phrase) to be similar to a given Model, by adopting M's

a) Behaviours; here, it is important to recall that imitation does not necessarily mean to learn a new set of "movements", but rather learn to give a new meaning (and also a new context) to a given behaviour. As Laland and Odling-Smee (1999: 6) observe, by referring to Heyes (1995), "... it is *not the motor pattern* that is learned, but rather existing *topographically defined behavioural elements*, alone or in combination, *are associated with the consequences of the behaviour*, in a particular context" (italics are mine). More explicitly, in imitation, agents learn to adapt their behaviour to achieve new goals.

b) Internal states, including the mental ones, such as beliefs, values, preferences (think of the dietary preferences among rats), goals, practical heuristics (think of the washing of sweet potatoes among Japanese macaques, Heyes & Galef, 1996). Internal states should not be confused with internal behaviours, that is, mental actions and operations, although these may also be targets of imitation.

c) Skills (think of Goodall's, 1964, well-known example of the skills necessary for foraging for termites using stalks acquired by infant chimps imitating adults).

d) External standards and criteria, which are inferred to (i) input A's behaviour, (ii) be mirrored in A's behaviour, and (iii) rule it.

The things that are imitated are (either learned cognitively or selected via biological evolution) relevant to agents' adaptation. Such relevance assumptions are essential if imitation (1) is to be at all possible and (2) will combine efficiency with effectiveness.

5.2.4 How to imitate

The mental process required by imitation is variable and ranges from the blind reaction of a baby duck following the first mobile object occurring in its perceptive field (which may happen to be an ethologist rather than its

mother) to a much more complex set of mental operations and representations.

In the case of imprinting, the mental properties required by imitation are rather poor, since the difficulties have been somehow managed at the evolutionary rather than at the individual level. In other words, the mechanism is not based upon by the single agent's mental representations nor allowed by its reasoning capacity. Rather, during the evolution of the species the evolutionary process has gradually selected a sensory-motor schema which allows the individual animal to answer adaptively some questions crucial for its own survival.

In most examples of imitation, and quite often among human adults, imitation requires a rather complex set of mental representations and processes:

a) Social beliefs, i.e. (i) information about M, his social status, mental states, etc.; (ii) information about M 's credibility, reliability, expertise, etc.. Imitation implies trusting M (or a set of agents, possibly coinciding with the whole social environment) as a source of information about adaptive behaviour (for an analysis of trust, see Castelfranchi & Falcone, 1998). However, the extent to which trust promotes imitation is variable.

b) Social reasoning, that is the capacity to infer M's goals, beliefs, values, etc. from his behaviours or appearance.

c) Relativized social goals, both the goal to acquire information about M and the goal to be similar to him, as long as he is believed to be a suitable model.

6 Advantages of the present analysis

Since imitation may be displayed even on the ground of built-in schemata and reactive behaviour, what is the use of a cognitive model as one presented in this paper? This question is even more crucial if one does not aim only at describing imitation among natural organisms, but also at implementing imitation in artificial systems: if there is a way to obtain the same result with low-complexity mechanisms (such as routines and production rules), why then bother with high-complexity, cognitive mechanisms?

There are several answers to this question, both at a scientific level and at the level of agent and multiagent systems applications.

6.1 To improve scientific understanding of social learning

It is yet unclear what can be learned via simpler mechanisms, to what extent social learning can be effectively achieved thanks to simpler mechanisms at the level of the agent. Certainly, a model of imitation which does not account for its cognitive ingredients will hardly enable us to distinguish social learning from pure social contagion. The main difference between these phenomena seems to reside precisely in the role played by the agents' mental processes in each of them: in social contagion, a given behaviour spreads automatically and easily, and often as quickly it decays. In social learning, modifications of the agent's states or behaviours is more robust and durable. The question is how such a difference can be explained and somehow reproduced.

Third, cognitive ingredients allow us to give more adequate and complete accounts of different forms of social learning, e.g., social facilitation and imitation. Indeed, a low-level definition of imitation as a mere behavioural phenomenon does not do justice to the ethological evidence that only animals like apes and dolphins do exhibit imitation, while many others exhibit only simpler types of social learning, such as social facilitation, if any at all. Why should this be the case if imitation were essentially based upon mechanisms such as matching between kinesthetic and visual images, enough elementary, or simple, to be executed by members of lower-level species?

Fourth, a cognitive model allows for an evolutionary, or at least a stepwise view of social learning and intelligence. It allows for different degrees and types of social influence to be investigated and some forerunners of social reasoning (reasoning upon others' minds) to be identified. For example, the capacity to use others as environmental bookmarks requires, and therefore gradually evolves into, the capacity to map the environment by deconstructing how others behave in it. As a consequence, certain forms of social facilitation may require as complex mental processes as those involved by imitation. But it is also the case that imitation represents an evolution of the processes involved in more elementary forms of social learning.

6.2 Socially intelligent agents for technological applications

Technological applications in the field of agents, require more sophisticated models of interactive and social competencies (for an argumentation of this claim, see Conte 1999b). In particular, the necessity to improve agents' capacity to learn from one another is largely shared by agent systems scientists. Attempts at implementing this capacity often draw upon classifier systems, adaptable agents, etc.. Two orders of questions arise here:

(a) How far can one go with the behavioural model of learning allowed by current solutions, such as classifier systems; on the other hand, what are the advantages for agent systems' applications of implementing intelligent social learning?

(b) More crucially, which properties are needed at the level of the agent to implement intelligent social learning?

6.2.1 Why implement intelligent social learning

Current learning systems are essentially stimuli-response systems, either symbolic (e.g. Learning

Classifier Systems, cf. Watkins, 1989) or sub-symbolic (e.g. neural nets).

Evolutionary Reinforcement Learning, Classifier Systems used for adaptive agents (Holland, 1992), allow the acquisition of new (social) beliefs, and the emergence of new strategies and agents (cf. Holland, 1995). Whilst these systems actually implement learning and evolving mechanisms, and have allowed to study the emergence and spread of interesting social phenomena, they do not yet allow to implement:

(a) The acquisition of attitudes, preferences, and other non-behavioural features, which implies that these be implemented at the level of the model, and, moreover, that they are recognised ad interpreted by the learning agents.

(b) Selective learning and resistance to change: how to implement at the level of the agent, given criteria for learning (learning what is desirable, fair, respectable, etc.)? This is essential to preserve some degree of system's robustness, and provide the agents with a relative capacity and criteria for resisting external, namely social influence. On the other hand, it allows to implement selective learning, and "desirable global effects" to emerge and spread. Selective social learning is essential to implement the spread of social norms and conventions in multiagent systems.

(c) Social models: these represent an interesting criterion of selective learning, and therefore an enforcement mechanism of conventions and social norms, in which given others (the so-called significant others) are assumed as good, convenient, reasonable, respectable, etc. models for imitation. As an additional advantage, to implement social models would promote the agent-based simulation study of the emergence of social hierarchies and structures (such as coalitions, alliances, etc.).

(d) Different attitudes towards learning: natural agents vary as to their capacity for and attitude to learning. To implement learning variety is essential to several domains of agent systems applications (believable agents, synthetic actors, multiagent systems, etc.), and requires a model of the processes and mechanisms which lead agents to *want* to learn.

6.2.2 How to implement intelligent social learning

To fulfil the tasks listed above, agents need to

(a) acquire social mental representations, that is social beliefs and social goals and intentions, including the goal to imitate others, as well as the capacity to

(b) attribute external and internal features to others, and update or instantiate models of others

(c) reason upon social beliefs, thereby generating new beliefs and take them into account while acting and imitating,

(d) form relativized social goals, that is social goals relative to social beliefs

(e) compare one's own knowledge base with that of others

(g) decide whether to imitate, solving potential conflicts goals among the goal to imitate and not to imitate, according to some criterion

(f) adopt external criteria for selective imitation (e.g., social desirability)

(g) decide which agents to imitate, instantiating social models to existing exemplars.

7 Summary

In section 2, some requirements of an adequate treatment of social learning were identified and found still wanting in the current models. The analysis presented in this paper seems to contribute to meet those requirements.

Both a core notion of social learning and some specific notions relative to the main processes leading to it - social facilitation and imitation - have been provided, which allowed both the similarities and the differences between these processes to be emphasised. This analysis presents two main characteristics:

a) Rather than focusing exclusively on emulation-based processes, and the improvement of one's performance, the *more general* phenomenon of one's acquisition of new information has been addressed.

b) Rather than grounding social learning on social reinforcement, social cognitive properties and mechanisms have been investigated, which seem to account for both the *similarities and the specificities* of the two phenomena of interest.

Finally, the utility of the approach presented here has been examined at both the scientific level and at the level of agent system applications.

References

A. Bandura, *Social Learning Theory*. General Learning Press, New York, 1971.

S. Blackmore, *The Meme Machine*. OUP, Oxford, 1999.

C. Castelfranchi, R. Falcone, Principles of Trust for Multi-Agent Systems: Cognitive Anatomy, Social Importance and Quantification, *Proceedings of the International Conference on Multi-Agent Systems (ICMAS 98)*, Paris, La Villette, July 1998.

P.R. Cohen, H.J. Levesque, *Persistence, Intention, and Commitment*, In P.R Cohen,J., Morgan,M.A. Pollack (eds), *Intentions in Communication*., 33-71. Cambridge, MA: MIT Press, 1990.

R. Conte, Social Intelligence Among Autonomous Agents, *Computational and Mathematical Organization Theory*, forthcoming, 1999.

R. Conte, C. Castelfrachi, *Cognitive and Social Action*. UCL Press, London, 1995.

J.L. Freedman, D. Perlick, Crowding, contagion and laughter. *Journal of Experimental Psychology*, 15:295-303, 1979.

B.G. Jr. Galef, Social enhancement of food preferences in Norway rats: A brief review. In *Social Learning in Animals: the Roots of Culture*. C.M. Heyes, B.G. Jr Galef (eds), Academic Press, New York, 49-64, 1996.

J. Goodall, Tool using and aimed throwing in a community of free living chimpanzees. *Nature*, 201, 1264-1266, 1964.

C.M. Heyes, B.G. Jr. Galef (eds.) *Social Learning in Animals: the Roots of Culture*. Academic Press, New York, 1996.

C.M. Heyes, Imitation and flattery: A reply to Byrne & Tomasello. *Animal Behaviour* 50, 1421-24, 1995.

R.A. Hinde, J. Fisher, Further observations on the opening of milk bottles by birds. *British Birds* , 44, 393-396, 1951.

M.L. Hoffman, Altruistic behaviour an the parent-child relationship, *Journal of Personality and Social Psychology*, 31, 937-943, 1975.

K.N. Laland, J. Odling-Smee, The Evolution of the Meme. Paper presented at the Conference "Can Memes Account for Culture?", Cambridge, UK, 4-5 June, 1999.

B. Latané, J.M. Darley, *The unresponsive bystander: Why doesn't he help?* Appleton, New York, 1970.

D.A. Levy, P.R. Nail, Contagion: A theoretical and empirical review and reconceptualization. *Genetic, Social and General Psychology Monographs*, 119:235-183, 1993.

P. Marsden, . Memetics and Social Contagion: Two Sides of the Same Coin? Journal of Memetics - Evolutionary Models of Information Transmission,2, 1998.

http://www.cpm.mmu.ac.uk/jom-emit/1998/vol2/marsden_p.html

G. Marshall, (ed.) Concise Oxford Dictionary of Sociology. OUP, Oxford, 1994.

R.W. Mitchell, A comparative-developmental approach to understanding imitation. In P.P.G. Bateson, P.H. Klopfer (eds.), *Perspectives in ethology, vol. 7: Alternatives*, 183-215, Plenum Press, New York, 1987.

D.P. Phillips, The influence of suggestion on suicide: Substantive and theoretical implications of the Werther effect. *American Sociological Review*, 39:340-354, 1974.

D.P. Phillips,The impact of fictional television stories on U.S. adult fatalities: New evidence on the effect of the mass media on violence. American Journal of Sociology, 87:1340-1359, 1982.

D.P. Phillips, The impact of mass media violence on U.S. homicides. American Sociological Review, 48:560-568, 1983.

H. Plotkin, *Darwin Machines and the Nature of Knowledge*. Penguin, London, 1994.

A.S. Reber (ed.),*The Penguin Dictionary of Psychology* (2nd ed.). Penguin, London, 1995.

E.H. Ritter, D.S. Holmes, Behavioral contagion: Its occurrence as a function of differential restraint reduction. *Journal of Experimental Research in Personality*, 3:242-246, 1969.

J. Searle, *The Construction of Social Reality*, Penguin, London, 1995.

R. Tuomela, M. Bonniver-Tuomela, From Social Imitation to Teamwork. In G. Holmström-Hintikka, R. Tuomela (eds) *Contemporary Action Theory*. Vol. II. Kluwer, Amsterdam, 1-47, 1997 .

E. Visalberghi, D. Fragaszy, "Do monkeys ape?" Ten years after. *Paper presented at the Workshop on "Imitation in Animals and Artifacts"*, 1999.

L. Wheeler, Towards a theory of behavioural contagion. *Psychological Review*, 73:179-192, 1966.

Appendix

List A

"Black-out" effect, or restriction of the space of possible actions. Here, no social competence operates, but a high regularity, or convergence, in agents' (social) behaviour due to some central extraordinary event. No mutual influence is exercised by the agents undergoing this effect. Still, they converge on the same behaviour (as happens in explosion of the birth rate nine months after a real black-out) thanks to a severe restriction of feasible actions.

Direct exposition, or the "party-shower" effect[6]. After the 1997/98 repeated earth moves in Central Italy, people were reported to develop compulsory paranoid thoughts. The same can be expected to be reported by the Turkish or Taiwan population after the more recent earthquakes in those areas. As in the black-out effect, a major discontinuity had been introduced in their normal life by a non-ordinary event. But unlike the previous effect, in this case, the influence of this event on agents is determined by their perception and interpretation of the event, and by the consequent feeling of powerlessness. However, neither influence nor imitation are (necessarily) at stake: agents did not need to communicate to, nor observe, one another (although, in fact, they most certainly did) for their feelings and behaviours to spread over the whole group.

Behavioural "domino" effect. With this type of effect, we enter in a more interesting sub-area of phenomena, namely transmission (and possibly convergence) due the non-mental effect of agents' behaviour on, and through, one another. Consider the case in which, in social or public settings (for example, a crowded restaurant[7]), you are obliged to raise your voice otherwise your friends won't be able to hear you. Here, agents do not form any representation of the others nor of their behaviour. They simply raise their voice in order to be audible, thereby causing a corresponding continuous increase of noise[8].

List B

The social models' influence. The propagation of mental anorexia among young women in Western societies is often considered as a consequence of their exposition to the unhealthy aesthetic standard of the "slender type". Of course, this does not account for the intrinsic replication success of the aesthetic standard in question (which is a memetic effect), but accounts for the width of the phenomenon: young women are strongly and widely influenced by it because fashion models and top girls *are* skinny. (This belongs to the same category of phenomena observed by Phillips, 1982, 1983 in his studies on the impact of media on social violence).

Socially-based goal-activation. Consider Weber's famous example discussed by several authors (for one example, Tuomela & Bonniver-Tuomela, 1997): while walking in the street you realize that people around you have opened their umbrellas. You then almost certainly infer that it is raining, although your thick hair or wide hat prevented you from perceiving the first drops. This inference will activate a goal of yours, i.e. not to get wet. Once such a goal has been activated, the role of the input agents stops. You are able to find a solution on your own: if you have an umbrella (which is already stored in your knowledge base as a good means to avoid getting wet), you will probably follow the example of your neighbours. But if you were not so mindful as to get one, you may decide to hasten your pace, or stop at the next pastry shop, or finally change your mind and get back on your steps. In all these cases, your decisions are influenced by your interpretation of the perceived passengers, but only in the former you actually replicate their behaviours (opening the umbrella).

Elite-oriented conformity. In this case, agents are ruled by their goal to show the same taste and preferences as those shown by (significant) others. They will exhibit given tastes and standards as long as they believe that these are shared by their models. Interestingly, this is complementary to the Simmel effect, shown by agents who consider themselves as "élites": these have the goal of maintaining preferences as long as these are shared only by their affiliates. As soon as others will converge on the same preferences, in order to be perceived as affiliates to the élite, the elitarian agents will drop them and turn to other, more selective, ones; and the process will be re-initialised.

---

[6]This name is after Searle's (1995) example of the prompt flight of participants at an out-doors party at the first evidence of an incipient shower.

[7]This example was shown to me by my colleague Cristiano Castelfranchi.

[8]This is also known as the "arena" effect: if during the performance, people in the first rows stand up, those who are right behind are automatically induced to follow their behaviour, and so on and so forth until people occupying the farthest seats.

The "vulnerable position" effect. On the highway, if everybody exceeds the speed limit, you are obliged to break the rule in order not to be hit sooner or later from behind. Your behaviour is influenced by the frequential norm established by others. However, neither imitation nor any representation of the other is (necessarily) involved. This is a mere case of an emergent regularity (which results in violating a specific norm).

Automatic contagion of emotion expression. At a party, if one starts to yawn, s/he will most certainly be followed by many participants. If you happen to listen to some foreigners speaking in an incomprehensible language and to see them bursting into laughter, you can't help laughing as well. If asked why you were laughing, you won't be able to give any good reason; still the automatic impact of laughter is irresistible.

The group effect (or social impact). The famous Social Impact Theory (Latané & Darley, 1970) accounts for an interesting variant of the vulnerable position effect in groups of agents facing an emergency. To avoid an isolated and therefore "vulnerable position", each bystander waits for someone else to make the first move and provide help to the victim. As a consequence, no-one will provide the help required.

Emotional sharing. Consider the case of empathy (cf. Hoffman, 1975). In this phenomenon, emotion spreads thanks to a specific mental process. A beggar shows helplessness and even despair because he is helpless (he believes something like "How dreadful: I am helpless"). The empathic passenger will feel sad if she believes "How dreadful: he is helpless". However, thanks to the empathic mechanism (rather mysterious, indeed, in absence of some biological source of solidarity), the passenger shares (to some extent and for a short time) the emotion or feeling expressed by the beggar. Here, something new occurs: the passenger perceives the emotional state of the beggar and infers his/her more general (social) state: empathy is in fact based upon specified attributions. In fact, people do not share the feelings of those who are perceived as responsible for their mishaps. Only under given attributions, they come to share the feelings of the victim. The emotional sharing is therefore caused by an inferential process, by a reasoning applied to the mental and objective conditions of the victim. However, no imitation occurs yet.

The Archaeology of Artificial Societies

Jim Doran
Department of Computer Science
University of Essex
Colchester, UK, CO4 3SQ
doraj@essex.ac.uk

Abstract

Can archaeologists help software engineers unravel what has been happening in an artificial society of intelligent agents? We discuss the methods that archaeologists regularly use and how they relate to the properties of an artificial society and the problems faced in recovering its history. As part of the discussion, an abstract model of a typical artificial society is presented, the structure of the process of interpreting evidence is analysed, and the particular macro-social phenomenon of socio-cultural collapse is considered.

1 Introduction

Archaeologists ask: "What did those guys (humans) DO the last few millennia?"

Software Engineers (will) ask: "What did those guys (agents) DO the last few hours/days?"

In this short paper I want to explore an idea that might at first sight seem a little bizarre: that archaeologists may have something to teach those who have to debug, or just understand, what has been happening lately in some specific *artificial society of "intelligent" agents*. It may be timely and useful to study how archaeologists recover past human social processes, in the belief that this will help future software engineers recover the past histories of the artificial societies that they are tasked to manage. Obviously, this task of recovery will be especially important if the societies in question have been malfunctioning. And, just because we *are* addressing societies, we may assume that the process of recovery, often preparatory to taking some kind of remedial action, will typically be pitched primarily at the social rather than the individual or the code level.

It perhaps needs emphasising that recovery of history in this context is indeed going to be a problem. Even if, implausibly, we assume that comprehensive tracing/logs exist in an artificial society -- but the supporting infrastructure would collapse? -- finding out what has been going on and why will *not* be easy. The problems are (a) the sheer magnitude and complexity of the raw activity logs that must be worked through, including traces reflecting the internal processing of agents, and (b) the restricted and costly access to the raw logs in practice. These problems exist whether or not the process of scanning activity logs is largely automated. Think of the "pile of printout" generated by even 10 minutes of a multi-agent simulation involving many non-trivial agents, if the tracing of the simulation and the agents within it is at all detailed!

2 What do Archaeologists do?

At the heart of archaeology is a set of techniques for working back from surviving evidence to the social processes, located in time and space, which gave rise to it. Crucial is the fact that some objects and types of material (especially stone, bone and pottery) can survive for very long periods of time in the ground. Archaeological excavation therefore enables the recovery of past activities, but the process of interpretation of archaeological data must allow for partial and differential survival. Further, archaeologists can only sample a small part of what is now in the ground, and frequently find themselves excavating a small faction of the whole site which they know to be there. This most notably occurs in the context of "rescue archaeology" when an archaeological site is about to be destroyed by some kind of building work. Thus the process of acquiring archaeological data is always time-consuming and subject to biases and problems.

Over the last hundred years and more, archaeologists have evolved a meticulous methodology involving systematic excavation, recording (including of stratigraphy in the ground), handling and restoration of artefacts, and documentation. A feature of archaeological method is that excavation and recording must proceed without too much prior assumption of what is important. The most seemingly insignificant fragment may prove highly informative.

Archaeologists first seek to establish what is in the ground, then work back to what existed in antiquity, and how and why it got there. The process of archaeological interpretation may be regarded as addressing three levels:

- the raw excavation data -- what is found and its context
- the micro (human) level of interpretation -- the human activities (e.g. cooking, burial, flint working, slaughtering) that the raw data reflect
- the macro (social) level of interpretation -- the rise, stabilisation and collapse of complex societies, the migration of a populations, and so on.

Some of the main archaeological techniques in standard use are:

- The recording and interpretation of stratification during excavation. This enables the original interrelationship in time and space of different deposits to be inferred.
- Comparisons between artefacts (for example, stone tools) leading to typologies, seriations and hence relative chronologies.
- Interpretations (e.g. as graves/hearths/hunting camps/kilns) based on common sense, and on documented examples of modern societies which are judged similar to (some of) those of antiquity.
- Absolute dating. Some materials surviving from antiquity may be dated with some precision. Two of the most important techniques are carbon-14 dating and dendro-chronology.

For detailed examples of these technique and their use, refer to any archaeological textbook (e.g. Doran and Hodson, 1975; Renfrew and Bahn, 1996). A discussion of a specific and typical piece of archaeological reasoning, and of how it might be automated, may be found in Doran (1970).

3 Artificial Societies

We may here take an *artificial society* to comprise located software agents, fixed and mobile, which are heterogeneous and which inter-relate. An artificial society is continuously active. Agents may be simple or complex, reactive or deliberative. If, as we may assume, the society is open, then agents enter and leave the society unpredictably. Agents may also self-clone or otherwise reproduce. There are potentially complex tasks to be performed, perhaps involving "documents" or other items, which are created and deleted, passed from one agent to another, exchanged, imported and exported, transformed and combined. Between the agents there may well be permanent and semi-permanent relationships and commitments, including dominance. At the social level (see Jennings and Campos, 1997), co-operative groups, organisations, markets and other social structures may either be designed into the society or may be *emergent*.

One widely discussed and plausible vision of the "post-PC" era of computing (e.g. CPHC, 2000) is of a world in which computing is distributed and networked throughout the environment in which we live, and is incorporated in the most mundane objects such as refrigerators, cars, door locks and even clothing. As these "things start to think" (in the current phrase), it seems that the post-PC vision is inevitably one of massive agent societies. It seems unrealistic to assume any kind of central control of such a society. Rather the infrastructure and the agents that it supports will be subject to diverse origin, ownership, and management, but with network wide standards and conventions to ensure coherence.

3.1 The Problem

What should happen if some kind of recent failure of the artificial society's functioning is detected, for example, if the total amount of useful activity has suddenly declined? To understand what has gone wrong, what has happened in the past must be understood.

What kinds of activity log will a "massive" agent society leave? It is safe to assume that the logs will be partial and heterogeneous, and that a software engineer will have access only to part even of the existing logs, which will be costly to obtain. We might reasonably assume that there is available:

- some evidence of agent location and movement
- some evidence of inter-agent communications and what they are about

but that as with humans:

- there is *little or no* record of what has been happening *inside* the agents.

3.2 An Abstract Model of an Artificial Society

In order to formulate the recovery problem more precisely, we may usefully consider the following *abstract model* of an artificial society and its associated recovery problem:

> There is a (very large) *network* (graph) whose nodes we call *sites*[1] and whose (bi-directional) links we call *channels*. At any moment many *agents* and *items* are located at sites on the network.
>
> Agents and items (and messages, below) have unique identifiers.

[1] This use of the word "site" is intended to echo the standard use of the term by archaeologists to denote any location of archaeological interest.

Each agent belongs to a certain *class* of agents (e.g. has a single, external owner). Many agents may belong to the same class.

Agents obey unobservable (by the software engineer) *internal decision rules*, which vary from one agent to another.

From time to time agents spontaneously appear at and disappear from sites. Thus the society is *open*.

Some agents are fixed (located permanently at one site). Some are able to move between sites (via channels).

Agents exchange messages of various types (via channels).

Items are of a number of types.

Agents perform the following actions upon items:

- create or delete an item (if agent and item at same site)
- pass an item from one agent to another via a channel (or two agents exchange items)
- import or export particular types of item from particular sites in the network
- transform an item from one type to another (if agent and item at same site)
- combine two or more items of appropriate types into a single item of a different type (if agent and items at same site)

There exist *item combination constraints* (known to the software engineer), which determine the outcome of the agent actions to combine items.

At each site there is maintained a (recent) history of (the identifiers of) incoming and outgoing agents, incoming and outgoing items and their types, and incoming and outgoing messages and their types. The content of messages is not recorded. At a moment of termination there is available to the software engineer a subset of these site histories.

The software engineer may assume that the intended *function* of the society is that each class of agent should create and export certain specific types of item.

The software engineer's task is to reconstruct as much as possible of the activity in the society, at a suitable level of abstraction, in the period up to a moment of termination.

3.3 Comment on the Abstract Model

It should be apparent from the model specification that the agents in the society are essentially tasked to create complex items by the appropriate use of *combination actions*. It follows that organised co-operation between the agents in the society will much enhance their effectiveness. Understanding of the society may therefore focus upon looking for stable patterns of co-operation and, perhaps, their "collapse" (see section 6.1). In turn these patterns of co-operation may well be reflected in patterns of messages recoverable from the site histories available to the software engineer.

For example, suppose that a group of agents is able to use a version of the well-known contract net protocol to perform a specific item compounding task, that is, many items of various types must be subjected to a set of combination actions so that a particular type of item is generated. One agent initiates the task and delegates item retrieval and item combination tasks to others and this process of delegation is repeated recursively. One may predict that a *distinctive pattern of messages* will often be recoverable from the assumed activity logs.

The nature of the items within the model is deliberately left open here. Item combination may loosely be compared with the construction of, say, an automobile. But items in artificial societies are in reality more likely to be text documents of one kind or another.

4 From Archaeology to Artificial Societies

Can we now map what archaeologists do onto what software engineers must do to unravel the history of an artificial society? We consider some important correspondences one by one.

4.1 Targeted and "Rescue" Excavation

As stated earlier, archaeologists are frequently obliged to excavate sites on an opportunistic basis, prior to the sites' destruction. For artificial societies this corresponds to the likely limited availability of site activity logs and to time constraints on the recovery problem, perhaps imposed by site log deletion procedures. The archaeological use of formal statistical sampling may carry over to the agent domain.

4.2 Excavation Technique

Archaeological excavation corresponds in artificial societies to the recovery of detailed activity logs from particular network sites. Although it seems likely that these records will be easier to assemble and interrelate than are the stratigraphic and other contexts that archaeologists must deal with, this may not always be the case, and general archaeological requirements of meticulous study without undue prior assumptions will apply.

4.3 Artefacts, and Relative and Absolute Chronology

The notion of an individual artefact has at least two possible analogues in artificial societies. Firstly, it may perhaps be associated with individual log entries. However, in artificial societies we can probably assume the availability of absolute dates/times for log entries as the norm. This implies that much of the archaeological concern with establishing chronology via such techniques as stratigraphic analysis, and artefact seriation based upon typology will not arise.

Alternatively, and perhaps more persuasively, an association may be made with the "items" included in the abstract model of an artificial society presented earlier. Then the construction of typologies and seriations over sets of artefacts might arise as a means to the understanding of the patterns of co-operative "work" upon items that a group of agents have evolved between them.

4.4 Micro-Interpretations

As indicated earlier, archaeologists regularly interpret raw excavation data in terms of basic human activities such as cooking, burial of the dead, hunting and so on. The corresponding activities in artificial societies seem to be such micro level agent interactions as communication, negotiation, delegation and argumentation. To recognise instances of such interactions algorithmically seems quite feasible.

4.5 Macro-Interpretations

Encompassing micro-level basic human activities there are macro-level social phenomena. Archaeologists address these where they can. For example, they examine the existence of different types of society at particular locations and times (for example, centralised and/or ranked), the relationships (for example, trading) that may exist between different types of society, and the processes that appear to have led to changes, for example, migration or social collapse.

Similar macro-level phenomena are to be expected in artificial societies. Large communities of co-operating agents are to be expected and these may be recognised, and the complexity or otherwise of their internal structure assessed. But there will be at least one important difference. A particular agent community may well have a single external "owner" which directly or indirectly sets its "top goals" (as allowed for in the abstract model of Section 3 – the notion of agent classes). External ownership of agent communities will presumably constrain the macro-dynamics of the society in ways yet to be understood.

Some important indicators that archaeologists use to recognise macro-structure, for example developed societies, are:

- the complexity and size of sites (and their length of existence)
- evidence of trade links -- for example, artefacts all or part of which are remotely sourced.
- evidence of ranking and specialisation (for example, in graves)
- indicators of "civilisation", for example, writing and monumental architecture.

All of these indicators, except perhaps the last, can be giving meaning in the artificial society context. For example, the complexity of a particular agent community is measured by the proportion of its agents which can be shown to be specialised to a certain type of item processing i.e. disproportionate use of particular combination actions.

5 The Process of Interpretation

There are clearly certain similarities between archaeological interpretation and the interpretation of activity logs from artificial societies. These may be summarised by saying that both involve a set of *interpretation rules*, which must be used to recognise certain *entities* (e.g. hut, burial, auction, negotiation, migration). From an artificial intelligence perspective, the combination of rules and ontology (a "conceptual repertoire") may loosely be regarded as a *frame hierarchy* (sometimes called a *schema hierarchy*) in which each frame contains both a characterisation of its corresponding entity, and also procedures for that entity's recognition, and in which the relationship which structure the hierarchy is *a kind of*. Interpretation is then a process of heuristic instantiation of some the frames in the hierarchy. This concept of a frame hierarchy with attached rules has been explored and implemented in, for example, the classic expert system CENTAUR (Aikins, 1983) in which the concepts represented in the frame hierarchy were disease entities. A closely relevant archaeological example is the PALAMEDE system of Francfort (1990) which addresses the archaeology of proto-urban eastern civilisations in about the Third Millennium BC. PALAMEDE "simulates" archaeological interpretation to the point of the recovery of macro-social dynamics (see next section), specifically the evolution of urbanisation.

It must be stressed that although in different problem domains there is similarity in the form of the interpretative process and its reliance on a combination of interpretation rules and an ontology, the actual rules and the actual ontology will surely differ from domain to domain. Thus the rules by which an archaeologist recognises the existence of, say, a prehistoric hut from traces in the ground, are quite different from those we would (implicitly) use to recognise a hut in existence now, for example by looking at it and doing some visual processing. And, of course, archaeologists work with specialised concepts (e.g. a "Levallois point", a

"horizon") which do not exist in the everyday repertoire at all. The implication is that *recovering the history of an artificial society will also require the development of a conceptual repertoire and associated interpretation rules, conditional on the types of activity logs available for study.*

6 Understanding Social Dynamics

In the preceding section I suggested that the process of moving from evidence to interpretation involved rules of interpretation and an ontology that includes processes. But at the macro-social level the ontology and its associated processes, that is, the social dynamics, are not well understood in either human or artificial societies. Thus even with good information about low level activities the macro-dynamics are potentially very hard to recover and understand. It is therefore nor surprising that archaeologists tend to be cautious at this level. In particular, they rarely speculate about social trajectories that might in principle be quite possible, but which happen not to have occurred in prehistory. This means that social theory from the perspective of the prehistoric archaeologist is tied quite closely to archaeological record as it exists.

6.1 Socio-Cultural Collapse

As an example, consider the important and much studied example of a prehistoric macro-social phenomenon is *socio-cultural collapse* (Renfrew, 1979), which occurs when an established and complex society relatively suddenly disappears from the archaeological record or, at least, becomes sharply diminished in its complexity. There are many instances of this phenomenon in the archaeological record, of which perhaps the best known is the collapse of the Mayan society in Central America towards the end of the First Millennium AD. Socio-cultural collapse is particularly relevant here because were it to occur in an artificial society, it might be expected greatly to diminish the society's useful activity.

Archaeologists have identified many possible processes of collapse, some purely internal, some including one or more external factors. These have included not only such obvious candidates as invasion, disease, and climate change, but also more subtle "domino" effects impacting population centres, and negative feedback loops within the actions of the ruling elite.

Experiments with the EOS multi-agent system (Doran and Palmer, 1995) have suggested that two broad categories of collapse are:

> *Change in the environment of the society, such that its structure ceases to be functional.*
>
> *In a society where agents reproduce and "die", failure of the society successfully and continuously to reconstruct itself.*

Much more research is needed into trajectories of collapse in artificial agent societies, with the emphasis placed on identifying categories of possible trajectory, rather than merely modelling observed real-world instances.

6.2 Emergent Social Complexity

Those tasked to engineer effective artificial societies probably need to understand a wider range of macro-social phenomena than do archaeologists. Unfortunately almost everything remains to be understood, including much about the origins of emergent social complexity. In this regard we may speculate that (i) agent reproduction is important, and that (ii) so is the ability of an agent to "sell its own labour". Reproduction enables collective evolution and the emergence of agents with co-operative characteristics that may not be immediately predictable. By "selling its own labour", we mean that an agent, *in what it judges to be its own "top-goal" interests*, agrees a deal in which it makes a semi-permanent commitment to another agent's goal set[2]. Such an agent ceases to be fully autonomous and becomes, in effect, merely the occupier of a role (in one interpretation of that term). For example, an information seeking agent may, if it has the "authority" and ability, rationally choose to occupy a "role" in this sense in return for an information "feed". Thus, we may speculate, roles emerge and hence organisations as composites of roles.

But do the benefits of enabling such processes in the agents of an artificial society outweigh the danger that emergent phenomena will deflect the society from its intended function? Recent discussions of agent-based software engineering methodology (e.g. Jennings, 1999; Wooldridge, Jennings and Kinny, forthcoming) are cautious on this point and tend to assume that agent system designers will wish to exclude potentially uncontrollable emergent phenomena. If this view prevails then the problem of history recovering may be kept relatively simple.

7 Conclusions and Future Work

It seems clear that it is productive to compare archaeological methodology with software engineering methodology for artificial societies, and that there are two directions for further more detailed work. The first direction is to explore in greater detail the comparison between the standard archaeological process of data interpretation and that needed to interpret activity logs of agent societies. Particular topics are (i) the development of more precise abstract models of artificial so-

[2] For steps in this direction, see the Generalised Partial Global Planning co-ordination mechanisms of Decker and Lesser (1998).

cieties together with definite algorithms able to perform the recovery task for them, and (ii) a more detailed and insightful study of the relationships between the ontologies and interpretation rules corresponding to the two cases. It may be, for example, that the assumption I have made here that there will be no record available in artificial societies of the *internal* processing of agents, nor of the actual *content* of messages, will need to be revised.

The second direction for future work is to further investigate similarities in the macro-level dynamics of human and of artificial societies. Greater understanding seems possible and likely to impact both theoretical archaeology and the design of artificial societies and, indeed, to contribute to the development of general social science.

Acknowledgement

I am grateful to Omer F. Rana for encouraging me to pursue this line of thought.

References

J. S. Aikins. Prototypical Knowledge for Expert Systems. *Artificial Intelligence*, 20: 163-210, 1983.

CPHC. The Future Development of Computing in the United Kingdom. Workshop organised by the *Committee of Professors and Heads of Computing*, Manchester, January, 2000.

K. S. Decker and V. R. Lesser. Designing a Family of Coordination Mechanisms. *Readings in Agents* (eds. M. N. Huhns and M. P. Singh), Morgan Kaufmann. pp. 450-457, 1998.

J. E. Doran. Archaeological reasoning and machine reasoning. *Archaeologie et Calculateurs* (ed. J.C. Gardin) CNRS, Paris. pp 57-67. 1970.

J. E. Doran and F. R. Hodson. *Mathematics and Computers in Archaeology*. Edinburgh University Press. 1975.

J. E. Doran and M. Palmer. The EOS Project: integrating two models of Palaeolithic social change. *Artificial Societies: the Computer Simulation of Social Life* (eds. N Gilbert and R Conte), pp. 103-125. London: UCL Press. 1995.

H-P. Francfort. Modelling Interpretative Reasoning in Archaeology with the aid of Expert Systems: Consequences of a Critique of the foundations of Inferences. *Interpretation in the Humanities: Perspectives from Artificial Intelligence* (editors R. Ennals and J-C Gardin). LIR Report 71. The British Library. pp 101-129. 1990.

N. R. Jennings. Agent Based Computing: Perils and Promise. *Proc. 16th Int. Joint Conf. on Artificial Intelligence (IJCAI-99)*, Stockholm, Sweden. (Computers and Thought Award invited paper) 1429-1436. 1999.

N. R. Jennings and J. R. Campos. Towards a Social Level Characterisation of Socially Responsible Agents. *IEE Proceedings on Software Engineering*, 144 (1), 11-25. 1997.

A. C. Renfrew. Systems Collapse as Social Transformation: Catastrophe and Anastrophe in Early State Societies. *Transformations: Mathematical Approaches to Culture Change* (eds. A C Renfrew and K L Cooke) Academic Press. Pp. 481-506. 1979.

A. C. Renfrew and P. Bahn. *Archaeology: Theory, Methods and Practice*. Thames and Hudson. (2nd Edition). 1996.

M. Wooldridge, N. R. Jennings and D. Kinny. The Gaia Methodology for Agent-Oriented Analysis and Design. *Journal of Autonomous Agents and Multi-Agent Systems 3 (3)* (to appear) 2000.

Reverse Engineering of Societies - A Biological Perspective

Kerstin Dautenhahn
Adaptive Systems Research Group, Department of Computer Science
University of Hertfordshire, College Lane
Hatfield Herts AL10 9AB, United Kingdom
K.Dautenhahn@herts.ac.uk

Abstract

This paper reviews important concepts from biology, Artificial Life and Artificial Intelligence and relates them to research into synthesising societies. We distinguish between different types of animal and human societies and discuss the notion of social intelligence. Consequences of *social embeddedness* for modelling societies at different levels of social organisation and control are elaborated. We distinguish between simulation models of societies and the synthesis of artificial societies. We explain why the Artificial Life bottom-up approach is the most promising direction for reverse engineering of societies. The correspondence between synthesised societies and natural (human, animal) societies is investigated, presenting a hierarchy of synthesised societies with increasing indistinguishability between synthesised and human societies.

1 Artificial Life

> "Artificial Life is the study of man-made systems that exhibit behaviors characteristic of natural living systems. It complements the traditional biological sciences concerned with the analysis of living organisms by attempting to synthesize life-like behaviors within computers and other artificial media. By extending the empirical foundation upon which biology is based beyond the carbon-chain life that has evolved on Earth, Artificial Life can contribute to theoretical biology by locating life-as-we-know-it within the larger picture of life-as-it-could-be." ([Lan89])

The general method to build life-like artifacts is to use natural and artificial systems as part of a *comparative study*. On the one hand artificial systems serve as models of natural systems in order to investigate open questions in biology [TJ94], on the other hand natural systems can serve as models for the construction of artificial systems. For the latter we find many successful implementations as 'imitations' of sensorimotor behavior in animals (e.g. snake-like robots [Hir93], walking machines imitating stick-insects [CBC+94], fly-like robot vision systems [FPB91], LEGO robots showing cricket phonotaxis [Web95], ant navigation with an autonomous robot [MLP+98]). Whether one adopts a 'strong' (creating life) or 'weak' (modelling and simulating life) attitude towards Artificial Life, the 'products', in particular the physical (robotic) implementations of Artificial Life research, can have a quality of their own. Recent developments in synthetic pets (to give a few examples: Sony: *Aibo*, a robotic pet dog; Omron: *Tama*, a robotic cat dog; Cyberlife Technology: *Creatures*, software pets; Mindscape: *Virtual Petz*, virtual dogs, cats, and human babies) still show the technical limitations, in particular the robotic examples, but they point towards a scenario where such agents can exist side-by-side with us in our office environment, public places as well as private homes (see issues of believability, anthropomorphism etc. which support human's perception of artifacts as 'alive' discussed in [Dau98]).

2 Emergence

In Artificial Life systems the term emergence is used if any properties of a system (e.g. the behaviour of an agent) arise from the system's interactions with the environment. Emergence is then neither a property of the environment, nor the agent or its control system. Usually the term is used with respect to levels of organisation, where properties which the system exhibits on a level A emerge from non-linear interactions of components at the lower level B (including other systems of the same type, the environment, and components of the system). The issues whether emerging properties need to be *novel*, or are inherently *unpredictable* (from the analysis of interactions

at level B), are controversial.

Langton ([Lan89]) discusses emergence with respect to the genotype-phenotype distinction. In biology, the genotype is the genetic constitution (genome) of an organism, while the phenotype refers to the total appearance of an organism (including behaviour), determined by interaction during development between its genotype and the environment. Identical genotypes might result in different phenotypes (cf. identical twins are not totally identical in appearance and behaviour), and similar phenotypes might result from very different genotypes. Applied to machines, Langton introduced the terms generalised genotype (Gtype) and generalised phenotype (Ptype), see figure 1, a. As with biological organisms, the Ptype of a machine cannot be predicted from its genotype (unless Gtype, Ptype and environment are trivially simple). Likewise, the Gtype cannot be 'designed' for a particular Ptype. A particular Ptype can usually only be achieved by trial-and-error experimentation (e.g. within a experimentally driven incremental design methodology) and/or by using evolutionary techniques.

Artificial Life systems are usually multi component systems. Single components on any level of granularity can be studied, e.g. components can be rules, processes, behaviours, individuals. The bottom-up Artificial Life approach of synthesising systems is fundamentally different from the traditional top-down approach of Artificial Intelligence (AI), as well as different from the analytical approach in biology. Braitenberg's *law of uphill analysis and downhill invention* points this out [Bra84].

> "It is pleasurable and easy to create little machines that do certain tricks. It is also quite easy to observe the full repertoire of behavior of these machines – even if it goes beyond what we had originally planned, as it often does. But it is much more difficult to start from the outside and to try to guess internal structure just from the observation of behavior. It is actually impossible in theory to determine exactly what the hidden mechanism is without opening the box since there are always many different mechanisms which identical behavior. Quite apart from this, analysis is more difficult than invention in the sense in which, generally, induction takes more time to perform than deduction: in induction one has to search for the way, whereas in deduction one follows a straightforward path." [Bra84], p. 20.

Revealing the mechanisms underlying animal behaviour (let alone animal minds) is usually a long and difficult endeavour. To give an example: observing an animal walking, climbing, swimming reveals very little about the biological neural control structure generating this behaviour. Numerous different controllers could be programmed which could generate a particular locomotion pattern, e.g. distributed or hierarchical controllers. A successful method in biology is the *hypothetico-deductive* approach, generating a hypothesis which is precise enough to make predictions about the outcome in particular experimental setups. Experimental setups on walking behaviour usually involve *disturbing* (interrupting, manipulating) the system and measure how the system copes and return to its normal normal pattern (e.g. involving obstacles or even leg amputation in stick insects). The investigation of walking behaviour in stick insects is a concrete example of the success of this methodology ([Cru90]), and results were specific enough to allow the construction of a robotic model ([DKS+98]).

What does this mean with respect to animal societies? First of all, large-scale 'experimentation' with animal (in particular human) societies is difficult and in the case of human societies certainly not desirable. Also, animal societies are being influenced and controlled by a huge number of factors and parameters (see different levels of organisation and control in section 3.4). Thus, relating the effects observed after a local disturbance of the system to particular control parameters of the system is practically extremely difficult, if not impossible. A straightforward way is therefore, as Braitenberg[1] suggested on the level of the individual, to *synthesise* social systems, as discussed in the next section. Most commonly computational (rather than physical) models are used as models of societies. However, as we will later see, building artificial societies in this way might be pleasurable and (relatively) easy, but creating realistic models has its own difficulties.

3 Artificial Societies

3.1 Modelling Human Societies

Artificial Societies as computational models of human (present or historical) societies have increasingly gained attention in the social sciences. [CHT97] discuss the following potential contributions of computer simulations to the social sciences:

- to direct attention to the study of emerging behavioural patterns, structures and social order

[1] Please note that the Braitenberg vehicles are *Gedankenexperiments*, neither computational nor robotic implementations. However, Braitenberg's ideas on how to incrementally, in a bottom-up manner, increase the complexity of a vehicle's behaviour – as it appears to the external observer – has significantly influenced the development of agent controllers in simulations and robots.

(e.g. cooperation, coordination, institutions, markets, norms etc.)

- to overcome the difficulties of conventional analytical or empirical research methods and techniques to investigate social dynamics and test corresponding theories and models (e.g. world models, population dynamics, in general: change, evolution and complexity of social systems)
- to study decentralised and self-organised social phenomena in increasingly unpredictable and complex environments

Artificial Societies are usually understood as agent-based models or 'laboratories' of social processes in which "fundamental social structures and group behaviors emerge from the interaction of individuals operating in artificial environments under rules that place only bounded demands on each agent's information and computational capacity." [EA96], p. 4. The *Sugarscape model* described in [EA96] shows impressive examples of modelling migration patterns, economic networks, disease transmission and other social processes.

The Journal of Artificial Societies and Social Simulation (JASSS) gives many examples of how artificial societies can help studying social processes ranging from anthropology to economics.

Different software environments are available at present for individual-based modelling (as opposed to models based on mathematical equations) of societies, among the most widespread in the Artificial Life and Social Simulation Community is the *Swarm Simulation System* (http://www.swarm.org/).

3.2 Modelling Insect Societies: Self-Organisation and Stigmergy

The term 'societies' is generally applied both to human and other animal societies, including social insects. Social insects (e.g. termites, bees, ants) are very well studied and two important theoretical concepts are used to understand coordination in social insect societies, namely *self-organisation* and *stigmergy*. Our fascination of social insect societies is based on the fact that we observe many impressive results of coordination among individuals, rather than complex behaviour at the level of the individual (e.g. building of huge and complicated structures like termite mounds, cooperative transport, foraging behaviour which seems to 'optimally' exploit environmental resources and can adapt to changes dynamically, seemingly complex 'planning' mechanisms necessary for sorting behaviour, and many more). Recently, models of *swarm intelligence* and their applications to problems like combinatorial optimisation and routing in communications networks have been studied extensively (see [BDT99], [TB99]). The concept *stigmergy* was first developed by the French zoologist Pierre-Paul Grassé in order to understand the emergence of regulation and control in social insect societies. Stigmergy is a class of mechanisms mediating animal-animal interactions [TB99]. According to [BDT99] and [TB99] two of such mechanisms are *quantitative stigmergy and self-organised dynamics* and *qualitative stigmergy and self-assembling dynamics*. Generally, the behaviour of each insect can be described as a stimulus-response (S-R) sequence (even for solitary species). If animals to not distinguish between products of others's activities and their own activity, then individuals can respond to and interact through stimuli. This does not require direct communication between individuals, individuals 'communicate' indirectly, via the environment. In *quantitative stigmergy* stimuli in the S-R sequence different quantitatively. Pheromone fields and gradients are examples of using quantitative stigmergy, e.g. the construction of pillars by termites. Here, termite workers impregnate soil pellets with pheromone and the pellets are initially randomly deposited. The initial deposits and their diffusing pheromones increase the attractiveness of the deposit. Once the deposits reach a critical size, pillars or strips emerge through a positive feedback loop (the more pheromones a pillar emits, the more it becomes an attractor for more deposits).

In *qualitative stigmergy* we have a discrete set of stimuli types, i.e. during nest building wasps do not add new cells at random. Locations with already existing three adjacent walls are preferred. Thus, once particular structures are finished they serve as qualitatively distinct stimuli. This principle which we observe on the level of animal-animal interaction can also be observed in solitary insects like *Paralastor sp.* wasps building a mud funnel: once the animal completes a particular stage in the building process, the structure serves as a new stimulus and triggers different responses. Experimental manipulation of the structure and the resulting response of the animal confirms the S-R sequence underlying the behaviour.

The second concept important for understanding social insect societies is self-organisation, or "a set of dynamical mechanisms whereby structures appear at the global level of a system from interactions among its lower-level components. The rules specifying the interactions among the system's constituent units are executed on the basis of purely local information, without reference to the global pattern, which is an emergent property of the system rather than a property imposed upon the system by an external ordering influence" ([BDT99], p. 9). Not unsurprisingly one of the first very successful Artificial Life research projects studied the emergence of global patterns in ants and robots ([DGF+91], [TGGD91], [DTB92]),

and has presumably shaped the understanding of the concepts of emergence and self-organisation in Artificial Life as much as theoretical work did. Self-organisation has four basic ingredients [BDT99]:

- Positive feedback. Amplification through positive feedback can result in a 'snowball effect'. Pheromones can increase the attractiveness of particular locations, e.g. trail laying and trail following in some ants species is used in recruitment of a food source.
- Negative feedback. It counterbalances positive feedback and in this way helps stabilising the overall pattern. The exhaustion of food sources or the decay of pheromones are examples of negative feedback.
- Amplification of fluctuations. In order to find new solutions self-organisation relies on random walk, errors, random task-switching etc.
- Multiple Interactions. Individuals can make use of the results of their own as well as of others' activities, but generally a minimal density of (mutually tolerant) individuals is required.

In Artificial Life, the term *collective* behaviour is generally used for group behaviour which is strongly genetically determined and does not involve direct communication between individuals, while the term *cooperative* is used for group behaviour which requires communication ([McF94]). Social insect societies and models thereof are typical examples of collective behaviour. Despite the influence of genetic factors in social insect behaviour, one should not forget that insects are sophisticated and highly complex animals which react dynamically and efficiently to state changes in the environment, themselves, or the colony. Deborah M. Gordon characterises the organisation of work, specifically task allocation, in social insect colonies as follows: "Individuals constantly alter their task status in two ways: they switch from one task to another, or move between a resting state and the active execution of some tasks. It is clear that both intrinsic and extrinsic factors contribute to task allocation. Individuals vary in predisposition to participate in certain tasks, and the tendency to perform a particular task changes as the individual grows older. Moreover, these age-dependent predilections are strongly influenced by at least two types of external cues: actions of other individuals, and events in the colony's environment." ([Gor96], p. 122). Thus, the individual and social life of an individual member of a social insect society is very complex, and far from fully understood (let alone its neurobiology). Computational or robotic models of insects have always been crude simplifications of the animal's natural capabilities and behavioural (if not mental) capacities.

With respect to methodological issues, it is interesting to note that many results on social insect societies have been obtained with *perturbation* experiments, which in the case of insects is both experimentally practical and ethically less controversial than experiments with humans (cf. section 2).

3.3 Social Embeddedness

Artificial Life agents are said to be *situated* if they are surrounded by their environment and if their behaviour depends on on-line, real world sensor data which is used directly in a (usually behaviour-oriented) control architecture. Socially situated agents are therefore agents that perceive and react to other agents. In biology the term socially situated applies to both social insect societies, as well as human societies.

Bruce Edmonds (1999) defines the notion of *social embeddedness* as follows:

> "An agent is socially embedded in a collection of other agents to the extent that it is more appropriate to model that agent as part of the total system of agents and their interactions as opposed to modelling it as a single agent that is interaction with an essentially unitary environment." [Edm99].

A socially embedded agent needs to pay attention to other agents and their interactions individually. This definition was suggested for reasons of practicality with respect to constructing agent systems [ED98]. However, for human animals who have a primate mind which is specialised in predicting, manipulating and dealing with highly complex social dynamics (involving direct relationships as well as third-party relationships), and who possess language as an effective means of preserving group coherence, 'social grooming' ([Dun93]), and communicating about themselves and others in terms of stories [Dau99b], social embeddedness becomes a conceptual requirement for modelling human agents. Humans are not only dealing with very complex relationships but seem to have mental 'models' of themselves, others and the social world (the interested reader is referred to literature on theory of mind and mindreading, e.g. [Whi91]). Humans, different from ants, live in *individualised societies* (as do other species of birds and mammals). An increasingly complex social field and an increasing need to effectively communicate with each other were likely to be among the important constraints in the evolution of human minds. Following the widely accepted *Social Intelligence Hypothesis* (e.g. [WB88]), and the recently suggested *Narrative Intelligence Hypothesis* ([Dau99b]), there are two interesting aspects to human sociality: it served as an evolutionary constraint which led to an increase of brain size in primates, this in return led to

an increased capacity to further develop social complexity. Although it is still unknown why hominids needed or chose to live in social groups, this *feedback principle* soon led to the development of highly sophisticated levels of organisation and control and human societies.

3.4 Levels of Organisation and Control

The terms *anonymous* and *individualized* societies are used in biology in order to describe two different types of social organisation. Social insects are the most prominent example of anonymous societies where group members do not recognize each other as individuals but rather as group members. We do not observe bees or termites searching for missing members of their colony. Although individuals adopt specific roles in a colony they do not show individuality or 'personality' in the same way as e.g. puppies in the same litter show. The situation is quite different in individualized societies which primate societies belong among. Here we find complex recognition mechanisms of kin and group members. This gives rise to complex kinds of social interaction and the development of various forms of social relationships and networks. On the behavioural level long-lasting social bonding, attachment, alliances, dynamic (not genetically determined) hierarchies, social learning, development of traditions etc. are visible signs of individualized societies. In humans the evolution of language, culture and an elaborate cognitive system of mindreading and empathy are characteristics of human social intelligence in individualized societies ([Dau97]). As a consequence of the latter, humans are not only paying attention to other agents and their interactions individually, but they use their mental capacities to reason about other agents and social interactions.

It is at present unclear to what extend the social intelligence of members of other animal species, in particular very social species like monkeys and *Cetaceans*, is similar or different from our own. Culture as such is unlikely to be a unique feature to human societies, the acquisition of novel behaviours in what we might then call 'proto-cultures' can be observed in animals. To give an example: traditions have been observed among troops of Japanese macaque monkeys ([Huf96]): Japanese macaques showed several examples of the acquisition of innovative cultural behaviours, e.g. sweet potato washing and wheat-washing was invited in 1953 by a young female and subsequently spreading to older kin, siblings, and playmates, eventually to other members of the troop. Other observed cultural behaviours are fish eating (as many newly acquired food sources initially spreading from peripheral males to adult females, then from older to younger individuals), and stone handling or stone play (initially spreading only laterally among individuals of the same age). Subsequently all these behaviours were passed down from older to younger individuals in successive generations (*tradition phase*). These examples clear show the influence of social networks on the *transmission phase* of novel behaviour: the nature of the behaviour and social networks determine how the behaviours are initially transmitted, depending on who is likely to be together in a certain context and therefore is exposed to the novel behaviour. Innovative behaviours of the kind described here have been independently observed at different sites. Various factors have been discussed which influence cultural transmission: environmental factors, gender, and age, and other social and biological life history variables. For example, unlike potato or wheat washing, stone handling declines when individuals mature.

The striking similarity of cultural transmission of novel behaviour exhibited by Japanese macaque monkeys and what we call human culture, questions the uniqueness of human societies. Note, that this behaviour is observed in monkeys, which do not show complex forms of social learning like imitation, and do not seem to posses higher-level 'cognitive' capacities necessary for complex social forms of 'primate politics' shown by non-human apes and humans (cf. discussions on imitation, mirror-test, and theory-of-mind). However, monkeys are excellent social learners (using widely non-imitative forms of social learning, e.g. social enhancement). Reader and Laland (1999) therefore argue that the meme concept (usually treated as uniquely human, [Bla99]) can and should also be applied to cultural transmission among non-human animals. Animal societies can appear in various forms. Human societies, human culture and human minds reflect in many ways their evolutionary origin in animal societies, animal culture and animal minds.

In order to distinguish social behaviour in social insect (anonymous) societies from human (individualized) societies we previously proposed the following definition of *social intelligence* and *artificial social intelligence* which could be applied to human societies:

> Social intelligence is "the individual's capability to develop and manage relationships between individualized, autobiographic agents which, by means of communication, build up shared social interaction structures which help to integrate and manage the individual's basic ('selfish') interests in relationship to the interests of the social system at the next higher level. The term *artificial social intelligence* is then an instantiation of social intelligence in artifacts." [Dau99a], p. 130.

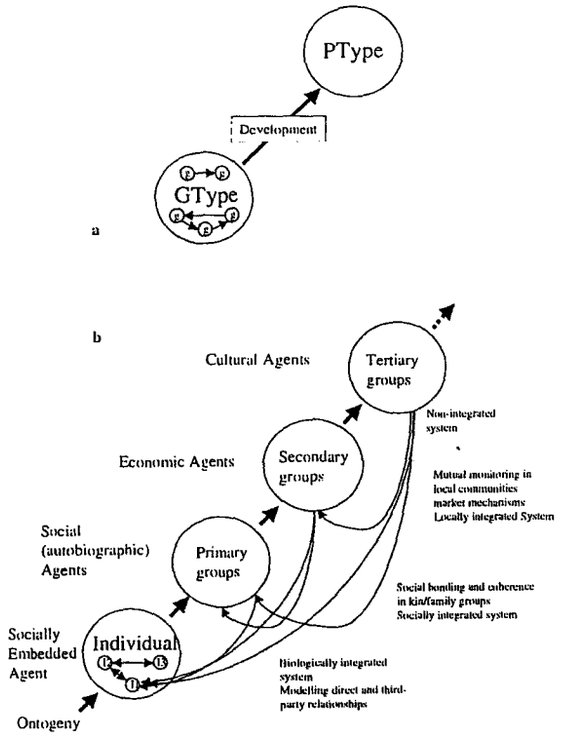

Figure 1: a) Emergence of behaviour in anonymous societies, b) emergence and feedback in individualised societies of socially embedded human agents on different levels of social organisation

This definition of social intelligence clearly applies to societies which are typical for highly individualized societies (e.g. parrots, whales, dolphins, primates), where *individuals* interact with each other, rather than members of an anonymous society. The definition therefore contrasts with notions of swarm intelligence and anonymous (e.g. social insect) societies (cf. section 3.2).

In [Dau99a] I suggested a hierarchy of different levels of social organisation and control, inspired by discussions on the development of social systems [HC95]. I distinguished between eusocial agents in anonymous societies where mechanisms of stigmergy and self-organisation (cf. section 3.2) result in a socially integrated systems[2], and human (individualized) societies where the individual is part of different levels of social organisation (primary groups, secondary groups, tertiary groups). The different 'roles' of a human as a individual, an autobiographic, social agent, an economic agents and a cultural agent are constraint by different mechanisms of social control.

What the hierarchical system of social organisation presented in [Dau99a] did not address sufficiently was the notion of social embeddedness as discussed in section 3.3. Considering that humans 1) have different roles and are socially situated on different levels of social organisation of control, and 2) are socially embedded in the sense that they can reason about themselves and their conspecifics, results in a sophisticated system of feedback and self-organisation among and between different levels of social organisation, as indicated in figure 1, b. The individual human and his/her behaviour on any of these levels is influenced by his/her knowledge about other levels, the levels cannot be clearly separated. Computational models of societies usually chose a particular level of granularity, e.g. modelling agents in kinship structures (primary groups according to the terminology above, e.g. [Tre95]), larger economic markets or settlements (comparable to secondary groups, e.g. [DP95], [BGPM+95], see also special issue on *computer simulation in anthropology* of JASSS, volume 2, issue 3, 1999), and cultural development and the evolution of memes (cf. tertiary groups, [Hal97]). Thus, in Braitenberg's words, simulating societies can be 'pleasurable', but the degree of 'easiness' depends on how faithfully we intend to model human beings as individuals, socially situated on different levels of social organisation, socially embedded in the sense that his/her behaviour is influenced by experiences and events on other levels of organisation. On an abstract level of modelling societies we might constrain agents to one particular level of granularity (and in this way avoiding feedback from other levels), and we could then observe effects of self-organisation resulting from positive and negative feedback, amplification of fluctuations and multiple interactions (cf. section 3.2). By introducing mechanisms of stigmergy we could even observe collective behaviour and global (temporal or spatial) patterns similar to those of social insect societies. But without modelling a socially embedded agent possessing social intelligence as defined above, we are unlikely to *synthesise* artificial societies rather than simulation models of (selected characteristics of) animal/human societies. However, the more elaborate computer simulations of societies become, the more we tend to label them as *artificial societies*. What evaluation criteria are useful in order to characterise the similarity between *real societies* and *artificial societies*?

In order to shed some light on the notion of 'simulating societies' versus 'synthesising artificial societies' we turn towards an issue which has been long

[2]Note that African naked mole-rats, mammals, show a eusocial organisation similar to social insects, [SJA91]. Thus, the eusocial form of organisation has evolved independently in different taxa of animals.

discussed in AI ('revived' through Artificial Life) namely the problem of reverse bioengineering (how to synthesise intelligence/life rather than analysing intelligent/living systems).

4 Reverse Engineering

Reverse Engineering, distinguished from standard (forward) engineering, is a widely used approach in software engineering. The problem here is (in short) to understand and extract the design of computer programme code which is not written by yourself. Moving towards an area more related to animals (as physical systems), reverse engineering is also popular for understanding products in order to redesign/improve or copy them (information about the original design process might be lost or inaccessible). The general idea here is to start with the product (e.g. a clock, a video camera etc.) and then to work through the design process in the opposite direction and reveal design ideas that were used to produce a particular product [3]. Stages in reverse engineering are system level analysis (e.g. estimating system cost, predict how system might work), subsystem analysis (e.g. identifying individual systems and how they interact), and finally component analysis where physical principles of components are identified. One approach towards analysing products is to regard the system as a black box with input and output and to identify how a) power, b) material and c) information is transformed or preserved.

Is reverse engineering applicable to animals as well as to artifacts? No matter how different the forward processes for animals ('design' by natural evolution) and artifacts (design by a human designer, starting from a specification) are, can we apply the reverse process to both kind of systems? Can we identify criteria similar to power/material/information in Reverse Bioengineering? In Dennett's discussion of such questions ([Den94]), he sympathises with the view of biology as reverse engineering, since biology tries to understand biological systems, its subsystems and components, and how they interact and work together. However, he argues that the top-down process of reverse engineering of artificial systems used for software or hardware are not appropriate for reverse engineering of natural systems (reverse bioengineering). The bottom up methodology of Artificial Life and the study of emergent effects is Dennett's favoured methodology for reverse bioengineering. Deducing the internal machinery of a black box is far more difficult than deducing the internal machinery of a system you synthesised (cf. Braitenberg's law of uphill analysis and downhill invention in section 2).

Forward engineering of artificial systems usually tries to eliminate unforeseen and undesired side-effects, namely emergent properties of how components locally interact with each other and the environment. Reverse engineering of products can therefore be very successful by decomposing the system into a system-subsystem-component hierarchy with well-defined interactions between elements on different levels, and with well-defined functions of each of the elements with respect to the whole system. A biological system, e.g. a human being, is a functionally integrated system which from a descriptive point of view can be decomposed into cells, tissues, organs, body, but this does not account for numerous self-organising and emergent effects down to processes within each cell. Elements in a biological system can have different functions. In evolutionary terms functions can change, new elements can evolve, new interactions between elements can occur. Thus, single functional elements are very difficult to isolate, in living systems 'side-effects' often prevail over fixed functional design. Thus, according to Dennett ([Den94]) Artificial Life is the most promising approach toward reverse bioengineering.

What we said above about reverse engineering of biological systems does naturally extend to societies. Thus, using computer simulations as models in order to understand natural societies as *Reverse Socioengineering* is no more different from the use of Artificial Life models (in software or hardware) in order to understand the behaviour of an individual (animal). More and more researchers in the field of 'individual artificial life systems' have recognised the need to build *complete agents*. Single aspects of an animal can be identified and modelled separately in a system which is, apart from that single aspect, very different from the natural model. However, such systems have often shown to be very limited in their explanatory power with respect to the overall behaviour of the animal. Building complete agents therefore tries to integrate as many aspects of the life of a natural system in an artificial system. Also, complete agents might ultimately not only simulate an animal, and appear 'life-like', but might develop as alternative life-forms. Concerning societies, when would we tend to call a system a true *instance* of a *society* rather than a *simulation model*? With respect to similarities between natural and artificial systems, one of the most widely discussed issues in AI (and Cognitive Science) is the *Turing Test*, discussed in the next section.

[3] Many publications are available on reverse engineering of software, but very little about reverse engineering of physical systems. This paragraph is therefore strongly based on lecture notes kindly provided by William Harwin who is teaching reverse engineering in a course on mechatronics at University of Reading.

5 Turing Test and Turing Indistinguishability

In Alan Turing's discussion of the question 'Can machines think?' he described an 'imitation game' which later became known as the 'Turing Test' (TT). The original formulation in [Tur50] of the imitation game was as follows:

> "It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'."[Tur50]

In order to address the issue of machine intelligence, Turing then suggested a variation of this test, namely having a *machine* taking the part of A in this game. The new question is then whether the interrogator will "decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" [Tur50].

In subsequent years, the standard interpretation of the Turing Test is to consider the scenario of a human, a machine and an interrogator, and the question whether a machine could 'pass' the test by communicating (traditionally in written format, via typewriter or computer) with the interrogator indistinguishably from a human being. If, in a particular experimental setup over a limited period of time, the interrogator is not able to distinguish between the two candidates (machine and human) then the machine is said to have 'passed' the TT. The machine (computer programme) is then either passing or failing the TT. Note, that this scenario of text-based, symbolic communication, although not unrealistic (cf. pen-pals or email-pals), substantially simplifies the process of natural human-human communication.

Although the TT can be dismissed as a 'trick', in the context of Artificial Intelligence and intelligent machines, the TT can serve as an empirical criterion, setting the empirical goal to generate human-scale performance capacity [Har92]. In [Har00], [Har01] Stevan Harnad extends the original TT scenario and proposes a TT hierarchy in order to discuss several *degrees* of indistinguishability instead of a yes/no evaluation. Note that each level subsume the capacities shown at lower levels.

- t1: toy models of human total capacity
- T2: Total indistinguishability in symbolic ('pen-pal') performance capacity (see standard interpretation of TT)
- T3: Total indistinguishability in robotic (including symbolic) performance capacity
- T4: Total indistinguishability in neural (including robotic) properties
- T5: Total physical indistinguishability

t1 is according to Harnad [Har01] the level of toy models, showing particular, narrow fragments of human capacity. All presently existing artificial systems have to be classified as t1 models. T2 refers to the well-known standard interpretation of the TT, it means that the machine is with respect to symbolic performance (language) indistinguishable from a human being. Note however, that this is not limited to a particular test-period, the hierarchy refers to life-long performance. Systems at level T3 are indistinguishable from humans with respect to 'robotic' performance, they show the same external sensorimotor (robotic) functions, such systems can 'mingle' with humans without being detected as machines. Systems at level T4 are indistinguishable from humans down to internal microfunctions, i.e. they possess artifical neurons, neurotransmitters etc. made of synthetic material, but showing the same functions (thus allowing e.g. organ transplantations between humans and T4 systems). Finally, systems at level T5 have identifical microphysical properties, they are engineered out of real biological molecules, physically identical to our own.

I suggest that the TT hierarchy, developed as a conceptual construct facilitating discussions on the synthesis and test of machine intelligence similar to human intelligence, also provides a useful means to discuss the issue of synthesising societies. I focus in the following on human societies, but non-human animal societies are included as well. The discussions are based on what we said in section 3.4 about human beings as individuals socially embedded in a hierarchy of social organisation and control.

- St1: *toy models of human societies.* At present, most existing systems of artificial societies and social simulation show particular, specific aspects of human societies. None of the systems shows the full capacity of human societies.
- ST2: *Total indistinguishability in global dynamics.* Computational social systems in the not too far future may show properties very similar to (if not indinstinguishable) from human societies. In particular domains, systems at this level might succeed to abstract from the biological, individual properties of humans and describe their behaviour on higher levels of social

organisation and control, e.g. processes in economics and cultural transmission might closely resemble processes we observe in human societies. Such systems might be used effectively as 'laboratories' in order to understand processes in historical and present societies, or might be used for predictive purposes.

- ST3: *Artificial Societies.* Total indistinguishability in social performance capacity. Societies at this level have to account for the socially embedded, individual and *embodied* nature of human beings. It might be possible that 'embodiment' in the sense of structural coupling between agent and environment can be achieved without requiring physical (robotic) embodiment (see [Dau99a] and [QDNR99]). The performance capacity of artificial societies at this level is indistinguishable from real societies, although the specific ways how these systems interact / communicate with each other need not be similar to or compatible with human societies. However, these societies go beyond 'simulation models' of societies, they *truly are* artificial societies.
- ST4: *Societies of Socially Intelligent Agents.* Artificial Societies at this level possess *social intelligence* like human beings do. This includes cognitive processes in social understanding in all aspects required in human societies, e.g. 'theory of mind', empathy etc. Members of artificial societies at this level might merge with human society, even in a physical sense (e.g. if the embodied agents are robots on a T3 or higher level, see above). However, the agents need not be robotic, they might exist as computational agents, with different means of communicating and interacting with each other.
- ST5: *Societies of Minds.* Total indistinguishability of social intelligence. The way these synthesised societies perform is not only indistinguishable from human societies with respect to their external performance, they are also indistinguishable with respect to the internal dynamics of their social 'minds'. Means and mechanisms of verbal and non-verbal communication, social 'politics', friendship, grief, hatred, empathy etc. at the individual level, as well as the performance of the society as a whole, is at this stage indistinguishable from human societies. Members of such societies could exist in human societies without any detectable difference, i.e. they might possibly consult the same psychiatrist.

The list above could help clarifying issues of correspondance and similarity between synthetised and natural societies.

6 Conclusion

The field of using agent-based computer simulations in social sciences and Artificial Life is still very young. This paper reviewed concepts from biology, Artificial Life and Artificial Intelligence relevant to simulating or synthesising artificial societies. This might help 1) avoiding to 'invent the wheel twice', 2) viewing the field in the more global context of system analysis and synthesis.

References

[BDT99] Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm intelligence. From Natural to Artificial Systems.* Oxford University Press, New York, Oxford, 1999.

[BGPM+95] S. Bura, F. Guérin-Pace, H. Mathian, D. Pumain, and L. Sanders. Cities can be agents too: a model for the evolution of settlement systems. In Nigel Gilbert and Rosaria Conte, editors, *Artificial Societies. The Computer Simulation of Social Life.*, chapter 6, pages 86–102. UCL Press, 1995.

[Bla99] Susan Blackmore. *The Meme Machine.* Oxford University Press, 1999.

[Bra84] Valentin Braitenberg. *Vehicles: Experiments in Synthetic Psychology.* MIT Press, Cambridge, 1984.

[CBC+94] H. Cruse, Ch. Bartling, G. Cymbalyuk, J. Dean, and M. Dreifert. A neural net controller for a six-legged walking system. In P. Gaussier and J.-D. Nicoud, editors, *Proc. From Perception to Action Conference, Lausanne, Switzerland*, pages 55–65. IEEE Computer Society Press, 1994.

[CHT97] Rosaria Conte, Rainer Hegselmann, and Pietro Terna. Social simulation – a new disciplinary synthesis. In Rosaria Conte, Rainer Hegselmann, and Pietro Terna, editors, *Simulating Social Phenomena*, pages 1–17. Springer Verlag, 1997.

[Cru90] Holk Cruse. What mechanisms coordinate leg movement in walking arthropods? *Trends in Neurosciences*, 13(15-21), 1990.

[Dau97] Kerstin Dautenhahn. I could be you – the phenomenological dimension of

social understanding. *Cybernetics and Systems*, 25(8):417–453, 1997.

[Dau98] Kerstin Dautenhahn. The art of designing socially intelligent agents: science, fiction and the human in the loop. *Applied Artificial Intelligence Journal, Special Issue on Socially Intelligent Agents*, 12(7-8):573–617, 1998.

[Dau99a] Kerstin Dautenhahn. Embodiment and interaction in socially intelligent life-like agents. In C. L. Nehaniv, editor, *Computation for Metaphors, Analogy and Agents*, pages 102–142. Springer Lecture Notes in Artificial Intelligence, Volume 1562, 1999.

[Dau99b] Kerstin Dautenhahn. The lemur's tale - story-telling in primates and other socially intelligent agents. Proc. Narrative Intelligence, AAAI Fall Symposium 1999, AAAI Press, Technical Report FS-99-01, pp. 59-66, 1999.

[Den94] Daniel Dennett. Cognitive science as reverse engineering: Several meanings of 'top-down' and 'bottom-up'. In D. Prawitz, B. Skyrms, and D. Westerstahl, editors, *Logic, Methodology and Philosophy of Science IX*, pages 679–689. Elsevier Science, BV, Amsterdam, North-Holland, 1994.

[DGF+91] J. L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chrétien. The dynamics of collective sorting: robot-like ants and ant-like robots. In J. A. Meyer and S. W. Wilson, editors, *From Animals to Animats, Proc. of the First International Conference on simulation of adaptive behavior*, pages 356–363, 1991.

[DKS+98] J. Dean, T. Kindermann, J. Schmitz, M. Schumm, and H. Cruse. Control of walking in the stick insect: from behavior and physiology to modeling. *utonomous Robots*, 7:271–288, 1998.

[DP95] Jim Doran and Mike Palmer. The EOS project: integrating two models of Palaeolithic social change. In Nigel Gilbert and Rosaria Conte, editors, *Artificial Societies The Computer Simulation of Social Life.*, chapter 6, pages 103–125. UCL Press, 1995.

[DTB92] J. L. Deneubourg, G. Theraulaz, and R. Beckers. Swarm-made architectures. In F. J. Varela and P. Bourgine, editors, *Proc. First European Conference on Artificial Life*, 1992.

[Dun93] R. I. M. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16:681–735, 1993.

[EA96] Joshua M. Epstein and Robert Axtell. *Growing artificial societies.* MIT Press, Cambridge, MA and London, England, 1996.

[ED98] B. Edmonds and K. Dautenhahn. The contribution of society to the construction of individual intelligence. Techical Report CPM-98-42, Centre for Policy Modelling, Manchester Metropolitan University, UK, 1998.

[Edm99] Bruce Edmonds. Capturing social embeddedness: a constructivist approach. Adaptive Behavior, 7:3-4 in press, 1999.

[FPB91] N. Franceschini, J. M. Pichon, and C. Blanes. Real time visuomotor control: from flies to robots. In *Proc. of IEEE Fifth International Conference on Advanced Robotics*, 1991.

[Gor96] Deborah M. Gordon. The organization of work in social insect colonies. *Nature*, 380:121–124, 1996.

[Hal97] David Hales. Modelling meta-memes. In Rosaria Conte, Rainer Hegselmann, and Pietro Terna, editors, *Simulating Social Phenomena*, pages 365–384. Springer Verlag, 1997.

[Har92] Stevan Harnad. The turing test is not a trick: Turing indistinguishability is a scientific criterion. SIGART Bulletin 3(4) (October 1992) pp. 9 - 10, 1992.

[Har00] Stevan Harnad. Turing indistinguishability and the blind watchmaker. Fetzer, J. and Mulhauser, G. (eds.) *Evolving Conciousness*, John Benjamins, Amsterdam (in press), 2000.

[Har01] Stevan Harnad. Minds, machines and turing: The indistinguishability of indistinguishables. Journal of Logic, Language, and Information, Special Issue on Alan Turing and Artificial Intelligence, (in press), 2001.

[HC95] Francis Heylighen and Donald T. Campbell. Selection of organization at

the social level: obstacles and facilitators of metasystem transitions. *World Futures*, 45:181–212, 1995.

[Hir93] Shigeo Hirose. *Biologically inspired robots: snake-like locomotion and manipulators.* Oxford University Press, 1993.

[Huf96] Michael A. Huffman. Acquisition of innovative cultural behaviors in nonhuman primates: a case study of stone handling, a socially transmitted behaviour in japanese macaques. In Cecilia M. Heyes and Jr. Bennett G. Galef, editors, *Social learning in animals*, chapter 13, pages 267–289. Academic Press, 1996.

[Lan89] Christopher G. Langton. Artificial life. In C. G. Langton, editor, *Proc. of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems, Los Alamos, New Mexico, September 1987*, pages 1–47, 1989.

[McF94] David McFarland. Towards robot cooperation. In D. Cliff, P. Husbands, J.-A. Meyer, and S. W. Wilson, editors, *From Animals to Animats 3, Proc. of the Third International Conference on Simulation of Adaptive Behavior*, pages 440–444. IEEE Computer Society Press, 1994.

[MLP+98] R. M"oller, D. Labrinos, R. Pfeifer, T. Labhart, and R. Wehner. Modeling ant navigation with an autonomous agent. In R. Pfeifer, B. Blumberg, J.-A. Meyer, and S. W. Wilson, editors, *From Animals to Animats 5, Proc. of the Fourth International Conference on Simulation of Adaptive Behavior*, pages 185–194, 1998.

[QDNR99] T. Quick, K. Dautenhahn, C. Nehaniv, and G. Roberts. The essence of embodiment: A framework for understanding and exploiting structural coupling between system and environment. Proc. CASYS'99, Third International Conference on Computing Anticipatory Systems, HEC, Lige, Belgium, August 9 - 14, 1999.

[SJA91] Paul W. Sherman, Jennifer U.M. Jarvis, and Richard D. Alexander, editors. *The Biology of the Naked Mole-Rat.* Princeton University Press, Princeton, N.J, 1991.

[TB99] Guy Theraulaz and Eric Bonabeau. A brief history of stigmergy. *Artificial Life*, 5(2):97–116, 1999.

[TGGD91] G. Theraulaz, S. Goss, J. Gervet, and L. J. Deneubourg. Task differentiation in polistes wasp colonies: a model for self-organizing groups of robots. In J. A. Meyer and S. W. Wilson, editors, *From Animals to Animats, Proc. of the First International Conference on simulation of adaptive behavior*, pages 346–355, 1991.

[TJ94] C. Taylor and D. Jefferson. Artificial life as a tool for biological inquiry. *Artificial Life*, 1:1–13, 1994.

[Tre95] Jean Pierre Treuil. Emergence of kinship structures: a multi-agent approach. In Nigel Gilbert and Rosaria Conte, editors, *Artificial Societies. The Computer Simulation of Social Life.*, chapter 6, pages 59–85. UCL Press, 1995.

[Tur50] Alan Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[WB88] A. Whiten and R. W. Byrne. The machiavellian intelligence hypotheses: editorial. In R. W. Byrne and A. Whiten, editors, *Machiavellian intelligence*, chapter 1. Clarendon Press, 1988.

[Web95] Barbara Webb. Using robots to model animals: a cricket test. *Robotics and Autonomous Systems*, 16:117–134, 1995.

[Whi91] Andrew Whiten. Natural theories of mind. Basil Blackwell, 1991, 1991.

The Inconstructability of Artificial Intelligence by Design
– the necessary social development of an agent that can pass the Turing Test[1]

Bruce Edmonds
Centre for Policy Modelling,
Manchester Metropolitan University
Aytoun Building, Aytoun Street, Manchester M1 3GH, UK
Phone. +44 161 247 6479 Fax. +44 161 247 6802
http://www.cpm.mmu.ac.uk/~bruce

Abstract

The Turing Test, as originally specified, centres on the ability to perform a social role. The TT can be seen as a test of an ability to enter into normal human social dynamics. In this light it seems unlikely that such an entity can be wholly designed in an 'off-line' mode, but rather a considerable period of training *in situ* would be required. The argument that since *we* can pass the TT and our cognitive processes might be implemented as a TM that, in theory, an TM that could pass the TT could be built is attacked on the grounds that not all TMs are constructable in a planned way. This observation points towards the importance of developmental processes that include random elements (e.g. evolution), but in these cases it becomes problematic to call the result artificial. The conclusion is that we will not be able to be able to implement an intelligence using only a *design stance*, but rather such intelligence requires considerable *social* development. In this light the TT can be read as challenging conceptions of intelligence which are disconnected with a social environment.

1 Social Dynamics of the Turing Test

The elegance of the Turing Test comes from the fact that it is not a requirement upon the *mechanisms* needed to implement intelligence but on the ability to fulfil a *role*. In the language of biology, Turing specified the niche that intelligence must be able to occupy rather than the anatomy of the organism. The role that Turing chose was a social role – whether humans could relate to it in a way that was sufficiently similar to a human intelligence that they could mistake the two.

What is unclear from Turing's 1950 paper, is the length of time that was to be given to the test. It is clearly easier to fool people if you only have to interact with them in a single period of interaction. For example it might be possible to trick someone into thinking one was an expert on chess if one only met them once at a party, but far harder to maintain the pretence if one has to interact with the same person day after day. It is something in the longer-term development of the interaction between people that indicates their mental capabilities in a more reliable way than a single period of interaction. The deeper testing of that ability comes from the development of the interaction resulting from the new questions that arise from testing the previous responses against ones interaction with the rest of the world. The longer the period of interaction lasts and the greater the variety of contexts it can be judged against, the harder the test. To continue the party analogy, having talked about chess, one's attention might well be triggered by a chess article in the next day's newspaper which, in turn, might lead to more questioning of one's acquaintance.

The ability of entities to participate in a cognitive 'arms-race', where two or more entities try to 'out-think' each other seems to be an important part of intelligence. If we set a trap for a certain animal in exactly the same place and in the same manner day after day and that animal keeps getting trapped in it, then this can be taken as evidence of a lack of intelligence. On the other hand if one has to keep innovating one's trap and trapping techniques in order to catch the animal, then one would usually attribute to it some intelligence (e.g. a low cunning).

For the above reasons I will adopt a reading of the Turing Test, such that a candidate must pass muster over a reasonable period of time, punctuated by interaction with the rest of the world. To make this interpretation clear I will call this the "long-term Turing

[1] This is a version of a paper to appear in a forthcoming special issue of the *Journal of Logic, Language and Information* (JoLLI) on "Alan Turing and Artificial Intelligence" in 2001.

Test" (LTTT). The reason for doing this is merely to emphasise the interactive and developmental *social* aspects that are present in the test. I am emphasising the fact that the TT, as presented in Turing's paper is not merely a task that is widely accepted as requiring abstract problem-solving ability, so that a successful performance by an entity can cut short philosophical debate as to its adequacy. Rather that it requires the candidate entity to participate in the reflective and developmental aspects of *human* social intelligence, so that an imputation of its intelligence mirrors our imputation of each other's intelligence.

That the LTTT is a very difficult task to pass is obvious (we might ourselves fail it during periods of illness or distraction), but the source of its difficulty is not so obvious. In addition to the difficulty of implementing problem-solving, inductive, deductive and linguistic abilities, one also has to impart to a candidate a lot of background and contextual information about being human including: a credible past history, social conventions, a believable culture and even commonality in the architecture of the self. A lot of this information is not deducible from general principles but is specific to our species and our societies.

I wish to argue that it is far from certain that an *artificial* intelligence (as validated by the LTTT) could be deliberately constructed by us as a result of an intended plan. There is an argument against this position that I wish to deal with: there is the contention that a strong interpretation of the Church-Turing Hypothesis (CTH) to physical processes would imply that it is theoretically possible that we could be implemented as a Turing Machine (TM), and hence could be imitated sufficiently to pass the TT. I argue against this in section 2 by showing that not all TMs can be deliberately constructed. If we can't construct a TM that could pass the LTTT, the other possibility is that we could implement a TM with basic learning processes and let it learn all the rest of the required knowledge and abilities. I will argue that such an entity would not longer be *artificial* in the section after (section 3). This is another way of saying that an intelligence that can pass the LTTT will be due to its social development as much as its original design. I will then conclude with a plea to reconsider the social roots of intelligence in section 4.

2 Why we can't Design all TMs

Many others have argued against the validity of the CTH when interpreted onto physical processes. I will not do this[2]. What I *will do* is argue against the inevitability of being able to construct arbitrary TMs in a *deliberate* manner. To be precise what I claim is that, whatever our procedure of TM construction is, there will be some TMs that we can't construct or, alternatively, that any effective procedure for TM construction will be incomplete. This is a strong argument because it follows regardless of the status of the physical CTH.

The argument to show this is quite simple, it derives from the fact that the definition of a TM is *not* constructive – it is enough that a TM could exist, there is no requirement that it be *constructable*.

This can be demonstrated by considering a version of Turing's 'halting problem' (Turing, 1936). In this new version the general problem is parameterised by a number, n, to make the *limited halting problem*. This is the problem of deciding whether a TM of length[3] less than n, and input of length less than n will terminate (call this TM(n)). The definition of the limited halting problem ensures that for any particular n it is fully decidable (since it is a finite function $\{1,\ldots,n\}\times\{1,\ldots,n\}\rightarrow\{0,1\}$ which could be implemented as a simple look-up table).

However there is not a general and effective method of finding the TM(n) that corresponds to a given n. Thus *what ever method* (even with clever recursion, meta-level processing, thousands of special cases, combinations of different techniques etc.) we have for constructing TMs from specifications there will be an n for which we can not *construct* TM(n), even though TM(n) is itself computable. If this were not the case we would be able to use this method to solve the full halting problem by taking the maximum of the TM and input's length finding the corresponding TM(n), and then running it for the answer.

What this shows is that any deterministic method of program construction will have some limitations. What it does not rule out is that some method in combination with input from a random 'oracle' might succeed where the deterministic method failed. The above arguments now no longer hold, one can easily construct a program which randomly chooses a TM out of *all* the possibilities with a probability inversely proportional to the power of its length (using some suitable encoding into, say, binary) and this program could pick *any* TM. What one has lost in this transition is, of course, the assurance that the resulting TM is according to one's desire (WYGIWYS – what you get is what you specified). When one introduces random elements in the construction process one has (almost always) to check that the results conform to one's specification.

However, the TT (even the LTTT) is well suited to this purpose, because it is a *post-hoc* test. It specifies nothing about the construction process. One can

[2] My position is that there are reasons to suppose that any attempt to disprove the physical CTT are futile (Edmonds, 1996)

[3] This 'length' is the base 2 logarithm of the TM index in a suitable recursive enumeration of machines.

therefore imagine fixing some of the structure of an entity by design but developing the rest *in situ* as the result of learning or evolutionary processes with feedback in terms of the level of success at the test. Such a methodology points more towards the constructivist approaches of (Drescher, 1991, Riegler, 1992 and Vaario, 1994) rather than more traditional 'foundationalist' approaches in AI.

3 The Necessity of *in situ* Development

At the end of the previous section, I raised the possibility that an entity that embodied a mixture of designed elements and learning in situ (using a source of randomness), might be employed to produce an entity which could pass the LTTT. One can imagine the device undergoing a training in the ways of humans using the immersion method, i.e. left to learn and interact in the culture it has to master.

However, such a strategy, brings into question the *artificiality* of the entity that results. (by 'artificial' I mean the extent that the object can be understood in terms of its design by us – thus a genetically modified crop is artificial to the extent that its characteristics have been designed rather than found. Although we can say we constructed the entity before it was put into training, this may be far less true of the entity *after* training. To make this clearer, imagine if we constructed 'molecule-by-molecule' a human embryo and implanted it into a woman's womb so that it developed, was born and grew up in a fashion normal to humans. The result of this process (the adult human) would certainly pass the LTTT, and we would call it intelligent, but to what extent would it be *artificial*? We know that a significant proportion of human intelligence can be attributed to the environment anyway (Neisser et al., 1996) and we also know that a human that is not exposed to language at suitable age would almost certainly *not* pass the LTTT (Lane, 1976). Therefore the developmental process is at least critical to the resulting manifestation of human intelligence. In this case, we could not say that we had succeeded in creating a purely artificial intelligence (we would be on even weaker ground if we had not determined the construction of the original foetus ourselves but merely copied it from other cells).

The fact is, that if we evolved an entity to fit a niche (including that defined by the TT or LTTT), then there is a real sense that entity's intelligence would be grounded in that niche and not as a result of our design. It is not only trivial aspects that would be need to be acquired in situ. Many crucial aspects of the entity's intelligence would have to be derived from its situation if it was to have a chance of passing the LTTT. For example: the meaning of its symbols (Harnad, 1990), its social reality (Berger, 1966) and maybe even its 'self' (Burns and Engdahl, 1998) would need to have resulted from such a social and environmental grounding. Given the flexibility of the processes and its necessary ability to alter its own learning abilities, it is not clear that any of the original structure would survive. After all, we do not call our artefacts natural just because they were initiated in a natural process (i.e. our brains), so why *vice versa*?

This is not just an argument about the word 'artificial'. These arguments have implications for the production of intelligent agents in terms of the necessity of considerable in situ acculturation. It may also have a bearing on how alien an artificial intelligence would be, should it arise, for such an entity would necessarily be considerably adapted to its environment and so probably comprehensible to other intelligences inhabiting a similar niche.

4 The Social Nature of Intelligence

All this points to a deeper consequence of the adoption of the TT as the criterion for intelligence. The TT, as specified, is far more than a way to short-cut philosophical quibbling, for it implicates the social roots of the phenomena of intelligence. This is perhaps not very surprising given that common usage of the term 'intelligence' typically occurs in a social context, indicating the likely properties of certain interactions (as in the animal trapping example above).

This is some distance from the usual conception of intelligence that prevails in the field of Artificial Intelligence, which seems overly influenced by the analogy of the machine (particularly the Turing Machine). Intelligence seems to be frequently taken as the presence of certain *machinery* that allows the solution of certain *problems*. This is a much abstracted version of the original concept and, I would claim, a much impoverished one. Recent work has started to indicate that the social situation might be as important to the exhibition of intelligent behaviour as the physical situation (Edmonds and Dautenhahn, 1998).

This interpretation of intelligence is in contrast to others (e.g. French, 1989) who criticise the TT on the grounds that it is *only* a test for human intelligence. I am arguing that this *humanity* is an important aspect of a test for meaningful intelligence, because this intelligence is an aspect of and arises out of a social ability and the society that concerns us in a human one. Thus my position is similar to Dennett's 'intentional stance' (Dennett, 1987) in that I am characterising 'intelligence' as a property that it is useful to impute *onto* entities because it helps us predict and understand their behaviour. My analysis of the TT goes some way to support this. It is for those who wish to drastically abstract from this to explain what *they* mean by intelligence – in what way their conception is useful and

what domain their definition relates to (typically more abstract versions of intelligence are grounded in 'toy' problem domains).

It is nice to think that Turing's 1950 paper may come to influence academics back to considering the social roots of intelligence, and thus counter an effect of his other famous paper fourteen years earlier.

References

Berger, P. L., *The Social Construction of Reality*, Garden City, NY: Doubleday, 1966.

Burns, T. R., and Engdahl, E. E., The Social Construction of Consciousness, Part 2: Individual Selves, Self-awareness and Reflectivity, *Journal of Consciousness Studies*, **5**:166-184, 1998

Cutland, N., *Computability*. Cambridge: CUP, 1980.

Dennett, D. C., *The Intentional Stance*, Cambridge, MA: MIT Press, 1987.

Drescher, G. L., *Made-up Minds – A Constructivist Approach to Artificial Intelligence*. Cambridge, MA: MIT Press, 1991.

Edmonds, B., 'Pragmatic Holism', *Foundations of Science*, **4**:57-82, 1996.

Edmonds, B. and Dautenhahn, K., 1998, 'The Contribution of Society to the Construction of Individual Intelligence'. *Workshop on Socially Situated Intelligence*, SAB'98. Zurich, August 1998. <http://www.cpm.mmu.ac.uk/cpmrep42.html>

French, R. M., 1990, Subcognition and the Limits of the Turing Test, *Mind*, 99:53-65. Reprinted in P. Millican & A. Clark (eds.). *Machines and Thought: The Legacy of Alan Turing*, Oxford, UK: Clarendon Press (1996).

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

Lane, H., 1976, *The Wild Boy of Aveyron*. Cambridge, MA: Harvard University Press.

Neisser, U. et al., 1996, Intelligence: knowns and unknowns, *American psychologist*, 51:77-101.

Riegler, A., 1992, 'Constructivist Artificial Life and Beyond'. *Workshop on Autopoiesis and Perception*, Dublin City University, Aug. 1992.

Turing, A. M., 1936, 'On Computable Numbers, with an application to the Entscheidungsproblem'. *Proceedings of the London Mathematical Society*, 2 42:230-265.

Turing A. M., 1950, 'Computing Machinery and Intelligence'. *Mind*, 59:433-460.

Vaario, J., 1994, 'Artificial Life as Constructivist AI'. *Japanese Society of Instrument and Control Engineers*, 33:65-71.

RECOGNITION OF INVESTMENT OPPORTUNITIES AND GENERATION OF INVESTMENT CYCLES

Guido Fioretti

University of Rome "La Sapienza", Dipartimento di Economia Pubblica
University of Stuttgart, Institut für theoretische Physik and Institut für Sozialforschung
guido.fioretti@mailcity.com

Abstract

Innovations cause entrepreneurs' mental models not to hold, generating optimism when innovations open up new fields of activity and pessimism when investments in fields that used to be safe no longer yield the usual returns. The state of optimism or pessimism in the minds of entrepreneurs eventually propagates to the whole economy, triggering up- and downswings of aggregate investments.

1 Introduction

Investing means to have a vision of the future, to believe it with a force and with a determination that goes well beyond the kind of "rational choice" utility maximisation can describe, and to commit oneself to a project that requires a great deal of will and persistence. When an individual risks his assets to become an entrepreneur, or when a manager risks his position in order to convince his company to pursue a certain strategy, there is much more involved than comparing the utilities of given alternatives. The "more" is a vision of these alternatives, of their possible consequences, and of a net of causal relationships that connect alternatives to consequences.

Many great economists stressed that investments are the engine of capitalism: Schumpeter (1911) first of all, and also Keynes (1936), who coined the expression "animal spirits" to designate the conviction by which entrepreneurs follow their visions. Unfortunately, formalisation of animal spirits has never been attempted, since it is an issue that is clearly out of the reach of the tools of decision theory.

This paper stems from a (possibly entrepreneurial) conviction: that formalisation of "animal spirits" is possible, and that it passes through the injection of some elementary concepts of cognition sciences into decision theory. An entrepreneur's "vision" is his mental model, a net of causal relationships connecting to one another mental categories that are peculiar to him in a way that is peculiar to him, a vision that the others can call "animal spirits" only if they don't share it, because their minds organise information in different ways. Consequently, if an entrepreneur's mental model is assumed to be known, it is possible to think sensible rules whereby this mental model is validated or rejected by empirical experience, and "animal spirits" arise or decay.

This is what this paper sets out to do. It does not present a realistic model, in the sense that one cannot use it (as it is) to describe a manager's behaviour. It is a methodological paper, in the sense that it proposes a new method. Clearly, this can be done best in an over-simplified setting by means of unrealistic assumptions, which have the purpose of isolating behaviour from influences by any other factor.

The paper consists of two main sections: the first one explains a single entrepreneur's decision model, the second one tests it in an artificial economy where many entrepreneurs interact with one another, as well as with consumers. Finally, a conclusion points to strengths and weaknesses of the above model, suggesting directions for future research.

2 Entrepreneurs' cognitive processes

Individuals simplify the mess of information they receive by classifying it into a manageable number of mental categories. Neither the number of mental categories, nor the criteria by which information is classified are constant with time; furthermore, the same individual may use different mental categories in different situations, and

categories may not be constructed around a single prototype (Lakoff, 1987; Clark, 1993).

However, for the sake of simplicity we shall neglect these insights: the mental categories we shall describe will be fixed in number, and they will obey fixed classification criteria. Nonetheless, they will retain the main concept of a mental category, namely that of a "box" whose content changes according to the novelties an individual encounters.

A mental model is a sort of map that provides orientation in decision-making by telling an individual at any time what it is 'normal' for him to expect; it is, in other words, a set of causal relationships that link a set of possible causes with a set of possible effects. A mental models can be seen as a net of connections that link mental categories to one another; however, the mental model that we shall ascribe to our decision-makers will be no more complex than a set of one-to-one relationships.

Let us describe entrepreneurs' cognitive processes by means of two kinds of mental categories: one for the 'actions' they undertake in order to produce and sell goods (usually entailing an act of spending), the other for the 'results' they obtain from customers (generally associated with proceeds). The 'situation' faced by an entrepreneur is a set of action-result pairs that occurred recently, while his 'behaviour' is the particular action he undertakes.

The actions entrepreneurs undertake, as well as the results they obtain, involve innovation of the qualitative features of goods, tastes and technologies, and can change with time in unpredictable ways. Nonetheless, we shall assume that entrepreneurs classify any action and any result in the following mental categories:

A_-: "Stand-by": the category for all entrepreneurial actions that require little money outlays, involve little innovation, but also little risk.

A_+: "Investments": the category of actions that, on the contrary, need large money outlays and involve important innovations, although they inevitably bear larger risks.

R_-: The category of mediocre results one normally expects from actions of category A_-.

R_+: The category of good results entrepreneurs expect from actions of category A_+.

Entrepreneurs evaluate the possibilities they are able to contemplate by means of a function u defined over their mental categories; let us stipulate that u measures utility if it is applied to result categories, and disutility if it is applied to action categories. Thus, it is obviously $u(A_-) < u(A_+)$ and $u(R_-) < u(R_+)$. Furthermore, since the prospect of a result of category R_+ must be such that it is convenient to undertake an action of category A_+ although it requires a larger effort than an action of category A_-, it must be also $u(R_+) - u(A_+) > u(R_-) - u(A_-)$.

If no innovation is introduced in the economy, the content of mental categories does not change with time; hence, in this case utility can be thought to be defined over the objects contained in mental categories. In this particular case, which is the one that is usually assumed, utility refers to goods, not to the way a decision-maker perceives goods.

If no novelty is supposed to appear, decision-makers only need to observe and measure the conditional probabilities of possible results for any given action, and take a decision that maximises their expected utility:

$$J(A) = \left[p(R_-|A)\,u(R_-) + p(R_+|A)\,u(R_+) \right] - u(A) \qquad (1)$$

where $A \in \{A_-, A_+\}$.

However, if innovations do take place, (1) no longer suffices. An innovation is recognised by a decision-maker by the fact that his old mental model does not work anymore. For example, a new technology may make typing machines obsolete and disclose new possibilities for computers; consequently, a mental model entailing causal relationships like "If I produce typing machines I will make good profits" and "If I produce computers nobody will buy them, because they are too big" may no longer be a reliable guide to decision-making. Thus, we must add to expected utility a term that measures the extent to which innovations cause a mental model to fail.

Let us assume that the mental model is constituted by the one-to-one relationships between action categories and result categories shown in fig.(1a): once an action of kind A_- has been undertaken a result of kind R_- is considered to be 'normal', and once an action of kind A_+ has been undertaken a result of kind R_+ is considered to be 'normal'. In other terms, the mental model of fig.(1a) says that good revenues are the normal outcome of investing, and that mediocre revenues are the normal outcome of hiding money under a mattress.

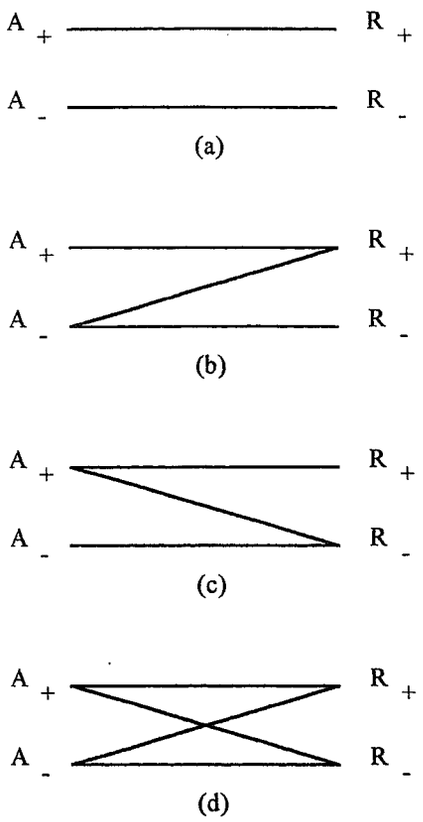

Figure 1

When correspondences other than those of case (a) occur, our entrepreneur may think that the categories and the model he is using are no longer appropriate to detect the relevant features of a reality where novelties are emerging. For instance, a new technology can open up new profit possibilities, a circumstance which would show up as the connections of case (b). But it can also cause the unexpected failure of investments on a field that used to be safe, in which case connections are as in (c). It can also produce both effects at a time, as it is shown in (d).

If the relations between mental categories are either as in (b), or (c), or (d), our entrepreneur is likely to think that this is the signal that an innovation is emerging, and that this innovation might have profound consequences for his activity. Independently of the probability distribution of successes and failures he might have computed in the past, misfunctioning of his mental model is likely to change the confidence he attaches to this probability distribution. Whether optimism as in (b), pessimism as in (c), or confusion as in (d), misfunctioning of his mental model inevitably alters his state of mind.

However, it is important to remark that a mental model says what it is 'normal' to happen, not what must happen all the times. Consequently, it is by no means obvious that relationships like those illustrated in (b), (c), (d) must be interpreted as failures of the mental model. They can also be interpreted as situations that occur rarely, though sometimes do: even with no innovation there is a positive probability that "safe" investments fail.

The trouble is that the emergence of novel profit possibilities shows up exactly in the same way as long-standing low-probability events do: it is up to the entrepreneur to decide whether what he is observing is the signal that the world is changing, or to discard it as worthless information. In other words, it is up to any single entrepreneur to detect the new profit possibilities opened up by an innovation, or to miss them. And from an entrepreneur's point of view, this is no little difference.

From the modeller's point of view, however, it easy to overcome this problem. In fact, it is sensible to assume that entrepreneurs use only the most recent information to evaluate the appropriateness of their mental model, while they use all information at their disposal to evaluate probability distributions of successes and failures. Only the most recent data can regard emerging innovations; on the contrary, it makes sense to use new and old information alike in order to calculate a probability distribution which is useful precisely to the extent that the present situation is analogous to the past ones.

Thus, let us assume that entrepreneurs are endowed with a memory of length L, and that they use all of its data to calculate a probability distribution of their successes and failures. On the contrary, their confidence in this probability distribution derives from the most recent data only, say the ones in the first $M \ll L$ memory locations.

The first M memory locations must be scrutinised in order to detect any deviance from the mental model illustrated in fig.(1a): if connections occurred, that are not those of the mental model, then the decision-maker does not have absolute confidence in the probability distributions he calculates using data from all L memory locations. Note that, unlike probability, it does not matter how many times one of these connections occurred, but only whether it occurred or not. Think to an entrepreneur who realises that some new technology allows high profits with relatively little effort, a circumstance which shows up as a set of connections like those of fig.(1b): it does not matter whether he receives one or two news about this novel possibility, the crucial issue is realising that such a new profit possibility exists.

Let us for simplicity neglect any process of information diffusion by assuming that information is immediately

available to all entrepreneurs, and for free. Each time an entrepreneur undertakes an action and receives a result, a new pair action-result enters the memories of all entrepreneurs (and obviously, the oldest pair exits at the same time). In this way, all entrepreneurs rely on the same data for their calculations; let us arrange these data in a matrix **D** of elements (A, R). With N entrepreneurs, matrix **D** has L rows and N columns.

Let us adapt to our oversimplified example a formalism that has been described elsewhere for the general case (Fioretti, 1998, 2001). Let us represent the relationships entailed in the first M rows of matrix **D** by means of a simplicial complex K made by two simplices A_- and A_+ having R_- and R_+ as vertices. In case (a), simplex A_- is constituted by the single point R_- while simplex A_+ is constituted by the single point R_+: the two simplices have no point in common and no simplicial complex exists, since A_- and A_+ are not connected. In case (b) simplex A_+ is still constituted by the single point R_+, but simplex A_- is the segment between R_- and R_+: the two simplices have vertex R_+ in common, simplicial complex K is made of a segment and one of its extreme points. Case (c) is analogous, just exchange R_- and R_+. In case (d) both A_- and A_+ are segments between R_- and R_+: the two simplices have one edge in common and simplicial complex K is constituted by two overlapping segments.

Let us stipulate that the complexity of a set of isolated simplices is zero, while the complexity of a simplicial complex is given by the dimension of the common face between the two simplices, plus one. Then, in the four cases of fig. A the complexity of simplicial complex K is $c_a(K) = 0$, $c_b(K) = c_c(K) = 1$ and $c_d(K) = 2$, respectively.

Let us agree that simplices A_- and A_+ contribute to the complexity of the whole simplicial complex in proportion to their dimension. Hence, $c_a(A_-) = c_a(A_+) = 0$; $c_b(A_-) = 1$, $c_b(A_+) = 0$; $c_c(A_-) = 0$, $c_c(A_+) = 1$; $c_d(A_-) = c_d(A_+) = 1$.

If complexity measures how "complex" a situation appears to the decision-maker, "confidence" must be its opposite. Let us define a 'degree of confidence' as follows:

$$\omega(A) := 1 - \frac{c(A)}{c(K)} \qquad (2)$$

where $A \in \{A_-, A_+\}$ and where $\omega \in \{\omega_a, \omega_b, \omega_c, \omega_d\}$ and $c \in \{c_a, c_b, c_c, c_d\}$, respectively.

This degree of confidence takes values in the $[0, 1]$ interval. In our case, these values are: $\omega_a(A_-) = \omega_a(A_+) = 1$; $\omega_b(A_-) = 0$, $\omega_b(A_+) = 1$; $\omega_c(A_-) = 1$, $\omega_c(A_+) = 0$; $\omega_d(A_-) = \omega_d(A_+) = 1/2$.

The objective function entrepreneurs maximise must entail a traditional part that depends on the probability distributions of successes and failures, as well as a cognitive part that depends on the confidence they have in their mental model. Let us use expected utility to represent the traditional part, but let us multiply it by the degree of confidence in order to represent the cognitive part.

However, the degree of confidence should have opposite effects when it refers to results that are better than those foreseen by the mental model, from when it refers to results that are worse than those foreseen by the mental model. For example, case (b) of fig. A should generate optimism, while case (c) should generate pessimism. In the first case, when the degree of confidence is less than one the objective function should increase. In the second case, when the degree of confidence is less than one the objective function should decrease.

Let us take account of this effect by raising the degree of confidence to the following exponent:

$$k(A) := \frac{u(R) - u(A)}{\overline{\Delta u}} \qquad (3)$$

where $A \in \{A_-, A_+\}$, and where $\overline{\Delta u}$ is the average of all $u(R) - u(A)$ differences. In our case $\overline{\Delta u} = [u(R_+) - u(A_+) + u(R_-) - u(A_-)]/2$, and $k(A)$ is subjected to the following restrictions: $k_a(A_-) = k_a(A_+) = 1$; $k_b(A_-) < 1$ and $k_b(A_+) = 1$; $k_c(A_-) = 1$ and $k_c(A_+) > 1$; $k_d(A_-) < 1$ and $k_d(A_+) > 1$.

Hence, our new objective function is:

$$J(A) = [\omega(A)]^{k(A)} [\, p(R_- | A)\, u(R_-) + + p(R_+ | A)\, u(R_+)\,] - u(A) \qquad (1')$$

where $A \in \{A_-, A_+\}$.

The central body of this objective function is its traditional part: expected utility, which makes sense as long as there is little innovation. On the contrary, innovations produce sudden jumps of the degree of confidence and sharp changes in the behaviour of entrepreneurs: the resulting investments can propagate to the whole system and eventually generate a business upswing.

Possibly, the meaning of the degree of confidence defined above can be better understood by making a reference to classifier systems (Holland, 1975). Classifier systems model categories by means of strings of zeroes, ones, and "don't care" characters # - this is not quite the nature of mental categories, and it is not an assumption of this paper; however, it is useful as a first approximation. From time to time, classifier systems activate procedures to create new categories, by means of random mutation of string characters or random recombination of parts of the old strings. A question classifier systems do not address is: When do these procedures need to be activated ?

The answer given in this paper is: categories need to be changed when the mental model they support does not work any more. After an individual recognised that his mental model is no longer a reliable guide to decision-making, but before a new mental model is constructed, an individual is hardly capable of decision-making. Sudden jumps in behaviour caused by (mis)functioning of an individual's mental model are particularly likely to happen in environments where novelties are the norm, as e.g. in the case of investments involving new technologies.

3 Investments Cycles

Let us now apply the above ideas to a model of investments cycles. The continuous introduction of innovations is supposed to keep the economy away both from perfect competition and monopoly; thus, our framework is rather that of imperfect competition.

Let us consider a population of entrepreneurs that interact with one another, as well as with a population of consumers. No other economic agents exist.

In order to allow each entrepreneur to trade with any other entrepreneur as well as with any consumer, let us assume that: i) All goods can be either consumed or used as production factors; ii) At least one good can be exchanged with any other. In order to avoid any constraint from past decisions on current production possibilities, let us also assume that: iii) No capital goods exist; iv) No inventories exist; v) Labour contracts refer to one production period only.

Entrepreneurs undertake actions towards other entrepreneurs as well as towards final consumers, receiving corresponding results from both kinds of agents. Consumers are supposed to behave in a more passive way: they return results to entrepreneurs, but they do not undertake actions by themselves. The actions entrepreneurs undertake consist of organising the production and sale of goods that generally entail some novel qualitative features, require new tastes to be appreciated and new technologies to be produced. The results entrepreneurs receive concern the reception of these goods by the market.

The categories by which actions and results are classified are A_-, A_+, R_-, R_+; however, due to restrictions $(i) \div (v)$ these categories can capture product innovation only. Technological innovation has a very limited scope in this model, since no capital goods exist: the only "technological innovation" this model is able to capture is learning by doing of pure labour. Tastes innovation is also minimal, since consumers accept the innovations proposed by entrepreneurs but do not carry out innovations on their own.

Let the number of entrepreneurs be fixed to N, and let $x \in [0, 1]$ denote the proportion of entrepreneurs who invest, i.e. who undertake an action of category A_+. It is $x = 0$ if all entrepreneurs undertake actions of category A_- and $x = 1$ if all entrepreneurs undertake actions of category A_+.

As long as at $x = 0$ all returned results belong to category R_- and at $x = 1$ all returned results belong to category R_+, the entrepreneurs' mental model is confirmed: they either observe (A_-, R_-) or (A_+, R_+) all the times. In this sense $x = 0$ and $x = 1$, which can be identified with "recession" and "growth", respectively, are equilibrium points.

However, even at equilibrium entrepreneurs do innovate: at $x = 0$ they innovate because they hope to get a result of category R_+, and at $x = 1$ they innovate because they fear that their competitors' innovations may turn their result into one of category R_-. Innovations can

be unexpectedly successful, which is the case when at $x = 0$ a result of category R_+ obtains. Or they can be unexpectedly unsuccessful, which is the case when at $x = 1$ a result of category R_- obtains.

Thus, even when the economy is at one of the two equilibrium points ($x = 0$ or $x = 1$), an unexpected result may suddenly change the degree of confidence and eventually push the system towards the other equilibrium. This mechanism can easily produce an irregular cycle where the economy continuously jumps from one equilibrium to the other.

The exogenous inputs to our model are: [I] the probabilities of obtaining a certain result category by entrepreneurs who are willing to undertake an action of a certain category, and [II] the probabilities of obtaining a certain result category by consumers.

Let inputs [I] be: $p_E(R_+ | A_-; A_-) = \varepsilon_-$, $p_E(R_- | A_+; A_+) = \varepsilon_+$, $p_E(R_+ | A_+; A_-) = \delta_-$, $p_E(R_- | A_-; A_+) = \delta_+$, where e.g. ε_- is the probability that an entrepreneur who is willing to undertake an action of category A_- returns a result of category R_+ to somebody who undertakes an action of category A_- towards him. It is obviously $0 \le \varepsilon_-, \varepsilon_+, \delta_-, \delta_+ \le 1$, with the following additional qualifications:

a) Entrepreneurs who undertake actions of category A_+ tend to give results of category R_+ more often than entrepreneurs who undertake actions of category A_-. Thus, $\delta_- < 0.5$ and $\delta_+ < 0.5$.
b) It is very unlikely that the cross connection of case (b) in fig. A occurs when one meets an entrepreneur who undertakes A_-, and it is equally unlikely that the cross connection of case (c) occurs when one meets an entrepreneur who undertakes A_+. Thus, $\varepsilon_- \ll 1$ and $\varepsilon_+ \ll 1$.

Let inputs [II] be $p_C(R_- | A_-) = \beta$ and $p_C(R_+ | A_+) = \gamma$, with $0 \le \beta, \gamma \le 1$. It seems reasonable to assume $\beta \gg 0$ and $\gamma \gg 0$.

Let us assume that the number of consumers is the same as the number of entrepreneurs, so that $p(R|A) = 1/2\, p_E(R|A) + 1/2\, p_C(R|A)$.

Probabilities $p_E(R|A)$ can be calculated by weighting probabilities $p_E(R|A;A)$ with the fraction of entrepreneurs who are willing to undertake actions of kind A_- or A_+, respectively. Hence, probabilities $p(R|A)$ can be expressed as follows:

$$p(R_-|A_-) = \frac{1}{2}\left[(1-x)(1-\varepsilon_-) + x\,\delta_+\right] + \frac{1}{2}\beta$$

$$p(R_+|A_-) = \frac{1}{2}\left[(1-x)\,\varepsilon_- + x(1-\delta_+)\right] + \frac{1}{2}(1-\beta)$$

$$p(R_-|A_+) = \frac{1}{2}\left[(1-x)(1-\delta_-) + x\,\varepsilon_+\right] + \frac{1}{2}(1-\gamma)$$

$$p(R_+|A_+) = \frac{1}{2}\left[(1-x)\,\delta_- + x(1-\varepsilon_+)\right] + \frac{1}{2}\gamma$$

(4a) (4b) (4c) (4d)

Equations (4a ÷ d) express objective function $J(A)$ as a function of x; in their turn, entrepreneurs undertake an action of category A_- if $J(A_-)$ is greater than $J(A_+)$, and vice versa. In this way the model is closed, and simulations can be carried out; figure (2) illustrates the corresponding flow chart.

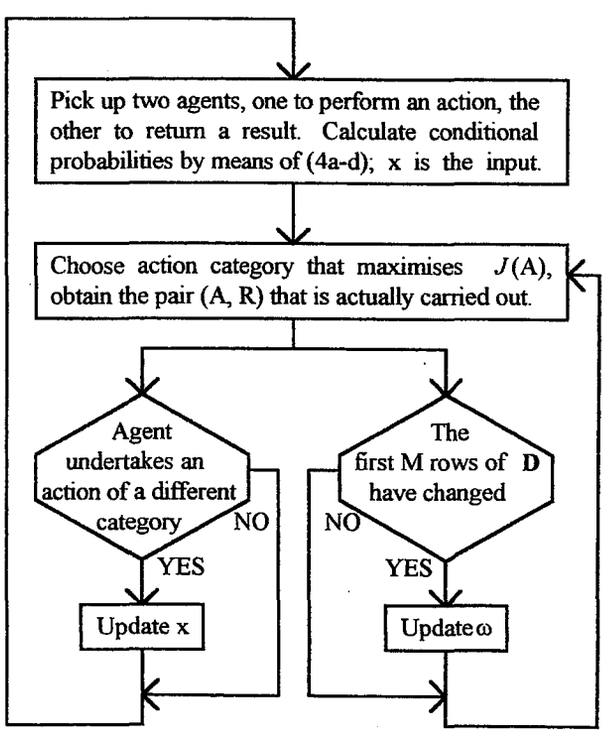

Figure 2

At each time interval one entrepreneur (it does not matter which one, since they all receive the same information) undertakes an action towards a randomly

chosen agent (who can be either a consumer or another entrepreneur), and receives a result from him. The $(A|R)$ pair obtained in this way enters matrix **D** and may produce a sudden jump of the degree of confidence.

At the same time, the action undertaken by the entrepreneur causes a smooth change of x and, through (4a ÷ d), a corresponding change of probabilities $p(R|A)$. In this way, at the beginning of the next time interval $J(A_-)$ and $J(A_+)$ take new values.

In order to compare simulations with one another, the random numbers generator will produce the same sequence of numbers throughout every run. Oscillations will be observed in x, the fraction of investing entrepreneurs. Since x must oscillate in $[0, 1]$, its initial value will be set at 0.5.

The following parameter set was chosen to illustrate the functioning of the model: N = 10; M = 5; $\delta_- = \delta_+ = 0.4$; $\varepsilon_- = \varepsilon_+ = 0.1$; $\beta = \gamma = 0.9$; $u(A_-) = 10$; $u(R_-) = 11$; $u(A_+) = 100$;

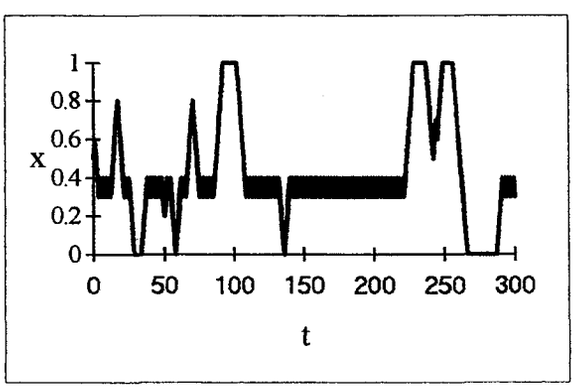

Figure 3

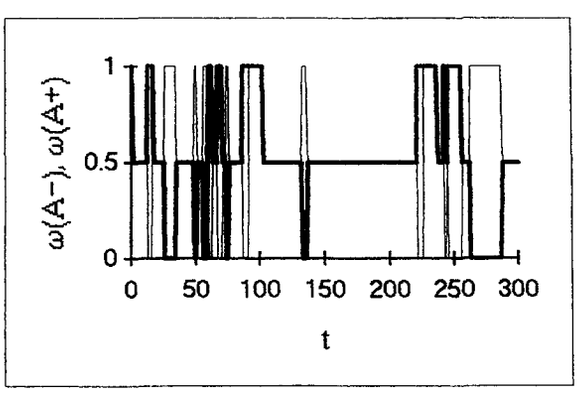

Figure 4

$u(R_+) = 100,000$. Fig.(2) depicts x, while fig.(3) shows the corresponding oscillations of $\omega(A_-)$ (thin line) and $\omega(A_+)$ (thick line). Each picture covers 300 runs.

Comparing figures (3) and (4) it is evident that, in general, periods of growth begin when $\omega(A_+)$ takes a high value and/or $\omega(A_-)$ takes a low value. The macroeconomic behaviour of the whole economy stems from microeconomic interactions, which on some occasions trigger avalanches that spread to the whole system.

Other simulations highlighted that the number of entrepreneurs N is a crucial parameter, since by increasing N oscillations become ever smaller. The reason is that the single action one entrepreneur undertakes causes x to vary by a $\Delta x = 1/N$; thus, if N is large trend inversions occur after x varied by smaller amounts. Possibly, this effect would disappear if the economy would be modelled as constituted by partial markets where a few entrepreneurs operate, instead of a single market every entrepreneur has access to.

Other important parameters are obviously L and M. In the simulation above it was assumed that probabilities $p(R|A)$, which entrepreneurs calculate from the data contained in their memories, could be expressed in terms of x: this amounts to assume that entrepreneurs know the state of the economy or, equivalently, that $L = \infty$. M was arbitrarily set to M = 5; simulations with M = 2 and M = 10 produced similar results.

4 Conclusions

The model presented in the above sections is likely to strike the reader for being a wear one, or perhaps - in the mind of a supportive reader - for being an innovative one. Let me try to assess its merits and limits, in order to help the reader to frame it within existing literature and to envisage the kind of research it might trigger.

Its basic merit, I believe, lies in modelling a piece of economic literature that has been deemed to be out of the reach of formal models hitherto. This should strike the reader as a possibility that could not be conceived before - a sudden link from a "A_-" to a "R_+" in the reader's mind, hopefully.

Also note that this is achieved without assuming the set of possible categories to be limited and known to the modeller, as e.g. when one assumes to represent mental categories by means of strings of given symbols and

given length. Clearly, this is the flip side of not being concerned with the way mental categories arise - the model above only deals with recognising that given mental categories are no longer appropriate.

The model makes a number of unrealistic assumptions, like e.g. absence of capital goods, perfect information, or the very fact of assuming that entrepreneurs have two action categories and two result categories only. Since this is a methodological model, and since the above limitations could be easily overcome in a more detailed setting, I do not think that this kind of critique really applies.

Rather, its basic weakness lies in its implicit assumption that the evaluation of a mental model's appropriateness can be described by means of an algorithm, as the procedure for computing the degree of confidence is. I actually do not think that this is true; to me, the algorithm presented above is just a useful representation in the researcher's mind.

The other weakness is an obvious consequence of not being concerned with the way mental categories and mental models arise: all entrepreneurs involved in the simulation have been assumed to be endowed with the same categories, since no procedure to describe how different categories interact and evolve has been designed. While this would be technically easy to do, I am not comfortable with assuming that the evolution of mental categories can be reproduced by simulations isolated from the real world Within the realm of algorithmic computation, the model presented in this paper already reached the limit of what it is sensible to do: mere detection of a mental model's misfunctioning..

Rather, I think that further research should endow agents with sensory organs that construct symbols and meaning at the same time, through the payoffs they receive from their environment (Cariani, 1998, 1998). This route is extremely difficult to pursue, because it implies that computer simulations are not enough - sensors that react to continuous, non-symbolic signals are needed. Nonetheless, I deem it is the only way to understand and reproduce the evolution of mental categories, and of the mental models that rest upon them.

References

P. Cariani. *Epistemic Autonomy through Adaptive Sensing*. Proceedings of the 1998 IEEE ISIC / CIRA / ISAS Joint Conference, Gathersburg (MD), 1998.

P. Cariani. Towards an Evolutionary Semiotics: The Emergence of New Sign-Functions in Organisms and Devices. *Evolutionary Systems*, ed. by G. Van de Vijver *et al.* Kluwer Academic Publishers, Dordrecht 1998.

A. Clark. *Associative Engines*. The MIT Press, Cambridge (MA), 1993.

G. Fioretti. A Concept of Complexity for the Social Sciences. *Revue Internationale de Systémique*, 12: 285-312, 1998.

G. Fioretti. A Subjective Measure of Complexity. *Advances in Complex Systems*, forthcoming, 2001.

J.H. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, 1975.

J.M. Keynes. *The General Theory of Employment, Interest and Money*. MacMillan, London, 1936.

G. Lakoff. *Women, Fire, and Dangerous Things*. The University of Chicago Press, Chicago, 1987.

J. Schumpeter. *Theorie der wirtschaftlichen Entwicklung*. Duncker & Humblot, Berlin, 1911.

About the Problem of Complexity and Emergence. The View of a Social Geographer

Dietrich Fliedner
(Formerly) Department of Geography, University Saarbruecken (Germany)
Email address: d.fliedner@rz.uni-sb.de

Abstract

Complexity, self organization, and emergence are difficult problems which have not yet been satisfactorily resolved. Systems and process research allow more far-reaching conclusions. There are 6 complexity levels:
1. solid bodies („solida") and „movements",
2. self ordering „equilibrium systems" and „movement projects",
3. self regulating „flow equilibrium systems" and „flow processes",
4. self organizing „non-equilibrium systems" and „conversion processes",
5. structurally self creating „hierarchy systems" and „hierarchical processes" (not treated here),
6. a materially self creating „universal system" and „universal process" (not treated here).
The autonomy of the systems increases with growing complexity.

In particular, it is necessary to distinguish flow equilibrium systems and non-equilibrium systems. The socalled complexity research did not sufficiently take this into consideration until now. Flow equilibrium systems distribute, and non-equilibrium systems convert information and energy. Non-equilibrium systems are limited by boundaries, consist of components which cooperate differently. Information and energy flows are separated and the elements have their specific tasks. The systems are internally vertically divided into 4 „bonding" levels which are characterized by the „main", the „task", the „control" resp. the „elementary processes". They are hierarchically ordered and must be performed completely, if the system is to fulfil its task of converting energy from one form in another as effectively as possible. This can be illustrated using an industrial company as an example. Further examples of non-equilibrium systems are perhaps biotic populations, organisms, cells, atoms, stars, etc.

The „emergence processes" intervene between the complexity levels and show the way from the simple to the complex. They always follow (geometrically) the same pattern, i.e. in accordance with a certain code:
1. *Bundling*: The processes of many elements [these are the solida or folded systems (see below, Folding) of the next lowest complexity level] are bundled in order to become components of the new process. In this way, the extent (amount) is fixed.
2. *Alignment*: These bundled processes are aligned to a comprehensive new process possessing 4 (or a multiple of 4) part processes. In this way, the number of process stages in this new complexity level is increased 4-fold in relation to the previous one.
3. *Interlacement:* The new process is now interwoven according to the new dimension constituting the complexity level concerned, i.e. the original basic orientation of the process, vertical or horizontal, is reversed by 90° in all its elements.
4. *Folding*:
In the final stage of the complexity process, one half of the process is folded behind the other in such a way that the beginning and the end of the process come into contact with one another. In this way, a limitation and possible control become possible, and the overall process continued.

1. About the current situation in research into complexity

How is order created from disorder, structure from non-structure and forms from amorphousness? In the past two decades, research into chaos and complexity have begun to look more closely at these questions. Many different disciplines have contributed to the work and astonishing results have been achieved. This in turn has given rise to the hope that true processes of self organisation might be simulated, i.e. that it might be possible to explain complex structures in the inorganic matter, the biotic world or the human society, or describe them in a reproducible way. However, there are signs that these hopes may have been premature. We are faced with a multiplicity of facts which are interpreted and evaluated in models in quite different senses. Attempts are being made to discover connections, but there is no unified theoretical basis. At any rate, all these experiments (e.g. "artificial society") have failed to produce explanations for self-organising and outlasting systems.

To date, the most power impulse for research into self-organising processes has come from the natural sciences, whereas the social sciences (in the broadest sense) were able to adopt the methods and results, albeit in a modified form. The question now is whether the social sciences in their turn can also provide new impulses for the discussion. The purpose of this poster is to attempt to indicate a new approach from a social-geographical standpoint. Information and energy, and their distribution and transformation play a decisive part in the concept pursued. First of all, it is necessary to define the

term system in its various forms. However, the processes should remain the central element of the examination, while giving due consideration to the fact that during their progress, processes do not only cause structures to be conserved or changed. The question now demanding an answer is: What tasks do the processes have with regard to the whole, the superior or overriding process? Because it is the examination of the development of the processes, the quality of which is highly differentiated, which enables a fresh approach to be taken to the problem of complexity and emergence, to the structuring of society and nature.

2. The various levels of complexity

Our reality can be depicted as an interwoven fabric of process sequences, which consists of various types of system, which in their turn, receive their substantial hold from carriers. The systems are linked to one another by energy flows which have to be insulated from one another to minimise noise or dissipation. Depending on how you look at one and the same object, you enter completely different levels of complexity. In all, 6 different levels can be identified which progress upwards from the simple ("solidum") to the highly complex (universum). These can be represented by specific process and system types. The tendency towards autonomy and self-preservation increases with the degree of complexity.

In their basic form, the processes consist of 4 stages. These lead from a first to a second state of the system. The process and system types which represent the individual levels of complexity are distinguished from one another by the degree of differentiation. In order to understand this, you have to examine which of the "system dimensions" are opened. Systems cannot be described with the usual geometric dimensions. Instead, it is necessary to take the links between the elements as a basis. In this way, you arrive at completely different dimensions. We must now distinguish between the dimensions of energy, time, hierarchy and space. With each level of complexity, a new dimension is included. In order to understand this, let us imagine a diagram in the form of a cross, in which hierarchy is in the top vertical, energy in the bottom vertical, time in the right-hand horizontal and space in the left-hand horizontal sections. In this way, the processes receive a certain basic orientation, i.e. vertical when dimensions of energy or hierarchy and horizontal when those of time and space are affected. For each of these 4 systemic dimensions, there are environments:
- energy dimension: the superior (energy demanding) and inferior (energy supplying) environment,
- time dimension: the preceding and succeeding environment,
- hierarchy dimension: the (hierarchically) superior and inferior environment, and
- space dimension: the spatial environment.

The structures of the first and second levels of complexity are not yet complex. Input and output are identical (solidum) or proportional to one another (equilibrium systems). The four other levels include self-control mechanisms, input and output are not proportional to one another, so that their systems may be described as complex. In order to compare the differences between the complexity levels with one another, their individual characteristics will be dealt with in sequence according to a fixed scheme:

System types:
- 1. complexity level: solidum (SOL)
- 2. complexity level: equilibrium system (ES)
- 3. complexity level: flow-equilibrium system (FES)
- 4. complexity level: non-equilibrium system (NES)
[5. complexity level: hierarchical system (not treated here)]
[6. complexity level: universal system (not treated here)].

Process types:
- SOL: simple movement (e.g. action "touching a plough");
- ES: movement project (e.g. action project "ploughing a field"). Many simple movements flow into one another. The system moves in conformity with the environment;
- FES: flow process, information and energy are distributed. From a structural point of view, several equilibrium systems are joined together. Conserving: demand for energy in wide sense (e.g. also goods) as information and supply of energy (or goods) keep one another in balance. Changing: increased or reduced demand for energy (it may contain the information on an innovation) spreads out from an initial location (diffusion) and changes the flow of energy. With stress, processes on the "edge of chaos". The energy dimension is optimized;
- NES (see the 4th section): work-division (= division of labour) process, transformation of information and energy. Many flow-equilibrium systems are linked with one another. In the "induction process", the system is oriented to the superior environment (see 2nd section, introduction) (e.g. the market) which demands energy of a certain quality, in the "reaction process", the system is dedicated to itself. Conserving: the supply of energy corresponds to the demand. Changing: when supply and demand do not correspond over a longer period of time, elements (= internally subordinated non-equilibrium systems) are added or taken away. The time dimension is optimised.

Carriers:
- SOL: every object which is moved, thereby transmitting energy. All concrete units, e.g.

individuals (bodies), limbs, perhaps artifacts, grains of sand etc.;
- ES: characteristic groups of every kind which are involved in movement projects, inasmuch as they are contiguous in space and time, e.g. the peasants in an agrarian community, perhaps the houses of a town, piles of sand etc. Elements of social systems are individuals to the extent that they carry out projects of movement (not individuals as such);
- FES: characteristic groups interacting with one another, all systems exchanging information and energy (in the sense of supply and demand), e.g. markets, interacting organisms (e.g. predator-prey relations), perhaps bow and strings of a violin, clouds in a current of air etc.
- NES (see the 4th section): differentiated work-division systems, e.g. social populations (families, companies, communities, city-umland populations, states), social or economic organisations, perhaps organisms, biotic populations, atoms, molecules, planetary systems, stars, galaxies, etc..

Energy transmission:
- SOL: direct transmission of force and impulse from the environment;
- ES: the elements and the individual movements adapt in the course of the movement project, with the result that the energy transmission from the environment to the system alters in the course of time;
- FES: differentiated transmission and distribution of energy. Demand by the superior environment (information flow), supply by the inferior environment (energy flow). Adaptation to the superior and to the inferior environment;
- NES (see the 4th section): transformation of information and energy in the induction process. The raw materials (= energy) from the inferior environment are brought together and changed into precisely fitting products. In the reaction process, the energy reaches the system itself.

Control:
- SOL: the movement of the solidum is controlled by the environment;
- ES: the movement of equilibrium system is controlled by the environments, internally by the neighbouring elements. The system orders itself;
- FES: by linking the end of the supply (= energy) flow with the beginning of the demand (= information) flow (by folding, see 3rd section) retroaction and therefore control of the flows becomes possible. The system regulates itself;
- NES (see the 4th section): the linking of the end of the induction process with its beginning (by folding) allows the control of the production height for the superior (demanding) environment (circle process). On this depends the shaping of the system in the reaction process. The process is controlled by the linking (work-division) of the information and energy flows in space and time. The system organises itself.

3. The emergence processes

Other processes intervene between the complexity levels. These represent emergence. This term includes processes which produce something new, which cannot be explained solely by the participating components. These emergence processes complete the picture of our reality as a fabric of processes consisting of many different levels and stages. The individual processes between the different complexity levels have a structure which is fundamentally the same, and can be described in the following terms:

Bundling:
The processes of many elements [these are the solida or folded systems (see below, Folding) of the next lowest complexity level] are bundled in order to become components of the new process. In this way, the extent (amount) is fixed. The processes of the elements are composed of 4 (or a multiple of 4, depending on complexity level) stages, and now serve the new process, i.e. the same purpose, each acting alone. In this stage of the emergence processes, they retain their fundamental orientation, vertical or horizontal (see the 2nd section, introduction). Within the co-ordinate system for each process, the process proceeds from the initial quadrant F(+x,+y) in a clockwise direction (vertically downwards and upwards) or in an anti-clockwise direction (horizontally, to the left and right).

Alignment:
These bundled processes are aligned to a comprehensive new process possessing 4 (or a multiple of 4) part processes. In this way, the number of process stages in this new complexity level is increased 4-fold in relation to the previous one. This new process has the same fundamental orientation, vertical or horizontal, like the individual processes of the elements as components. In this way, the new process has become a unit.

Interlacement:
The new process is now interwoven according to the new dimension constituting the complexity level concerned, i.e. the original basic orientation of the process, vertical or horizontal, is reversed by 90° in all its elements. In this operation, the position of the partial processes in the co-ordinate system must be observed. Now the new processes run at right angles to those of the next lowest or next highest complexity level, according to the new dimension being opened.

Folding:
In the final stage of the complexity process, one half of the process is folded behind the other in such a way that the beginning and the end of the process come into contact with one another. If the main

process is vertically oriented, the lower half is folded behind the upper one on a horizontal hinge (except in the case of the simple movement in the first complexity phase, representing the base). If it is horizontally oriented, the lower half is placed to the left of the upper one, and then folded behind it on a vertical hinge. In this way, a limitation and possible control become possible, and the overall process continued.

4. Model of the non-equilibrium system (see the 2nd section)

Returning to the problem outlined at the beginning (see the 1st section): up until now, it has not been possible to simulate outlasting non-equilibrium systems, e.g. a social population (or an atom ?). The course of the process itself, i.e. the sequence of the individual stages, is organised in the systems. The non-equilibrium system is the actual centre of activity of our reality. As already shown, information and energy are transformed in it, whereby the system preserves and organises itself.

Regarded vertically (energy dimension), the system occupies a position between the market demanding a certain product or the superior environment on the one hand, and the inferior environment supplying the necessary energy on the other. Internally, we distinguish four levels (assigned hierarchically to one another), which we may call "bonding levels". These organise the flow of information (i.e. here demand) moving vertically downwards and the flow of energy (i.e. supply) moving upwards. The processes proceed (horizontally) at the bonding levels. The processes of the bonding levels located lower down, supply those of the superior bonding levels. Each process level contains an entire four-part process ("basic process") of the inferior bonding level. Thus the time dimension is optimised, represented by a process sequence:

Main processes:
At the first bonding level, the "main process" is localised. The four "main stages" are as follows:
1. A stimulus (i.e. demand for the transformation of energy or matter) is received from outside, i.e. from the market:
"adoption".
2. The stimulus is implemented, the energy (or matter) is transformed: "production".
3. The stimulus is received to transform the structure of the system according to the new demands: "reception".
4. The stimulus is implemented, the system transformed: "reproduction".
The population is stimulated by the demand (at the beginning of the first main stage). This is met by its supply, which (as with the other stimulated systems of the same level) reaches the market at the end of the second main stage after a certain delay, since each process requires a certain amount of time. In the meantime, the demand has also changed. As this occurs repeatedly, oscillations are created.

Task processes:
At the second bonding level, the 4 main processes are further sub-divided and connected with one another in time. Here, contact is made with the preceding and succeeding environment. This means receiving the raw goods or raw material, and passing on the product in the induction process and conserving or possibly transforming the population itself in the reaction process.

Within the fixed framework of the system, this means that the activities differing in quality are joined together and thus the course of the process itself is structured. What should happen in the individual stages is established. In other words: the tasks of the processes making up these stages are of importance. Thus, we may speak of "task processes" and "task stages". In the context of (the main process) adoption:
1. Identification of the demand for a certain product, stimulus strength: "perception";
2. Decision whether the additional work can be taken on: "determination";
3. Distribution of the potential work to the elements (workers), which thus become adopters: internal diffusion ("regulation");
4a. The adopters, thus stimulated, may become producers: transmission of the stimulus to the second main stage, production, through spatial contacts: "organisation" (1st part).
In the course of (the main process) production:
4b. Receipt of the stimulus through spatial contacts from the first main stage: "organisation" (2nd part);
5. Distribution of energy to the elements (workers): "dynamisation";
6. Execution of the work: "kinetisation";
7. Supply of the demanded products to the market: "stabilisation".

Control processes:
The third bonding level is characterised by the linking of the system as a whole and the elements; here, the co-operation between the system (the population) as an entity ("system horizon") and the elements (the individuals, "element horizon") is controlled. This characterises the system-internal superposition. The process carried out with the elements is guided and controlled. For this reason, we use the term "control process" and "control stages". In the context of (the main process stage) adoption:
1st control stage: receipt of the stimulus (demand) equally by all the co-operating elements involved. The elements appear as homogeneous individuals;
2nd control stage: passive receipt of the stimulus according to system capacity. Stimulating (incoming) and receiving units (elements) form characteristic groups;

3rd control stage: receipt of the stimulus by the elements; supply of stimulus (work) and demand for it must be in flow equilibrium, i.e. (demand supplying) system and (demand demanding) elements form an internal flow-equilibrium system;
4th control stage: the elements become the (unified) element horizon as opposed to the system horizon. The system horizon A demands energy for transformation from the element horizon B, which is in contact with the inferior (energy supplying) environment. System horizon and element horizon depend on one another. The system is now a non-equilibrium system, i.e. the whole population becomes involved in the process.

Elementary processes:
This also applies for the 4th bonding level; here, the elements of the system (population) make spatial contact with the inferior (energy delivering) environment. The fact that all the processes demand space and that their effect on the environment varies in depth, has to be taken into consideration. In this sense, a spatial value can be assigned to each of the result values found in the various control stages (see above, Control processes). As the structure of the system changes, the space required also changes. As the demand increases (at the beginning of the adoption stage), new (so to speak) "space elements" are introduced, into a pre-determined volume. As the processes proceed within the context of the elements, they are termed "elementary processes" and "elementary stages".

4 Main, 16 task, 64 control and 256 elementary processes (but only 20 formulas) fully define the process structure of the non-equilibrium systems.

5. Significance of the above for research into complexity

In my view, we have generally reached a critical point in the development of our research in self-organisation which forces us to do much re-thinking. Chaos research and the related branches of the fields dealing with non linearity, non-equilibrium states, complexity, emergence etc. have developed models which operate on the basis of flow-equilibrium systems. They do not take sufficient account of the parts played by time and quality (tasks) within this context. It was scarcely possible for the natural sciences to identify the significance of quality for the processes. In this respect, the human society offers a wide enough range of phenomena and the stimulus to make these useful for research into complexity.

References:

For a more detailed treatment (and formalisation) of this summary, refer to the following books by the same author:

- Die komplexe Natur der Gesellschaft. Systeme, Prozesse, Hierarchien. Frankfurt a.M., New York etc. (Peter Lang) 1997;
- Komplexität und Emergenz in Gesellschaft und Natur. Typologie der Systeme und Prozesse. Frankfurt a.M., New York etc. (Peter Lang) 1999.

Also:

- Society in Space and Time. = Arbeiten aus dem Geographischen Institut der Universität des Saarlandes, Vol. 31. Saarbrücken 1981;
- Sozialgeographie. Berlin, New York (de Gruyter) 1993.

A New Look Into Garbage Cans - Petri Nets And Organisational Choice

Sven Heitsch, Daniela Hinck, and Marcel Martens

University of Hamburg,
Department for Computer Science and Institute of Sociology
Vogt-Kölln-Straße 30, 22527 Hamburg, Germany
{5heitsch, hinck, 5martens}@informatik.uni-hamburg.de

Abstract

Understanding how organisations make decisions is a crucial step towards understanding organisations. Seeing organisations as a place of structure and rationality led to unsatisfying results. The "Garbage Can Model of Organizational Choice" of Cohen, March, and Olsen (1972), fundamental to behaviouristic organisational theory, looks at "organized anarchies" and opens eyes for ambiguous and unpredictable decision situations. Reference Nets, a high-level Petri net formalism, offer formal semantics, graphical representation, means to model concurrency, and immediate executability, and, thus, seem to meet basic requirements to model and present sociological theories. In this paper Petri nets are used to formalise the Garbage Can Model and expose its implicit assumptions. The resulting model serves as a basis for interdisciplinary collaboration. Weaknesses of the original theory are laid open leading to new sociological considerations.

1 Introduction

Usually sociological theories are available as natural language texts and, thus, elude from formal analysis. To find clear semantics which is a prerequisite for formal analysis, verification of consistence, and executability, often is difficult. This paper reports on approaching a sociological model of organisational decision making with means of Petri net theory. In the socionics project at the University of Hamburg the emphasis lies in the modelling and analysis of sociological scenarios, aiming at evaluation and improvement of different theories. Both, for advancement in sociology and for better understanding of artificial societies (also see Sozionik@UHH, 2000).

Our chosen example of a sociological theory, namely the "Garbage Can Model of Organizational Choice", deals with decision making processes in organisations. And the way Cohen, March and Olsen (1972) do that, marks a point of changing the common view to such processes. This change of view refers to the *context* and the *order* - or better: the absence of order - in decision making processes. Here promising points for the actual research on organisations are touched which will be discussed later in the paper.

The "Garbage Can Model" is a fundamental and often cited contribution to behaviouristic organisation theory. The model combines empirical characteristics, theory, and simulational aspects. It also deals with the essential sociological task how organisations can survive while struggling with ambiguous and complex problems just as an unpredictable environment. The Garbage Can Model turns away from the common view that organisations are the right place for rational, intentional and well structured decision making. Rather there are seemingly a lot of incoherent actions and the results are not as intentional and desirable as they might be. The issue, whether this interpretation is a grounded one or a question of perspective, will be taken up later in this paper. At least, it is argued by the authors, that parts of any organisation can be described with this model at various times.

Originally, C. A. Petri (1962) intended to introduce a universal formalism for complex systems, offering formal semantics, explicit means to model concurrency, graphical representation, and executability. Elementary Petri nets consist of three static elements: places and transitions which are connected by arcs. Anonymous tokens represent the dynamic aspects by being moved from one place to another through switching transitions.

The high-level paradigm of "nets in nets" by Valk (1987, 1998) allows the tokens to be Petri nets themselves. This idea is incorporated and extended in Reference Nets by Kummer (1998). Each Petri net can be seen as an object (or even agent) in a Petri net environment.

This paper is based on a case study approaching the sociological theory (CMO, 1972) with Reference Nets. The emphasis is in the construction of an executable model which serves as a starting point for interdisciplinary collaboration and the validation and evaluation of the sociological theory. The Petri net model delivers new insights to strengths and weaknesses of the original contribution about organisational decision making. It provides a base point for connecting reflections which are new to the sociological discourse.

Other studies which are regarding the Garbage Can Model in a computational way have focused on artificial intelligence and simulational aspects (see Masuch and LaPotin, 1989).

The following section introduces the basic concepts of the Garbage Can Model of Organizational Choice. Section 3 gives a brief overview on the Reference Nets which are used as the modelling technique of the nets of section 4. In Section 5 the implications and results of this work are discussed. The last section concludes the paper and takes an outlook on relevant topics in the near future.

2 The "Garbage Can Model of Organizational Choice"

This section introduces the Garbage Can Model of Organizational Choice by Cohen, March, and Olsen (1972). Then a generalised version of the original work is presented. This will be the basis for the executable object Reference Net model of section 4.

The "Garbage Can Model of Organizational Choice" (1972) still is a relevant contribution to organisation theory because of the remaining actuality and applicability for present organisational processes. The authors led various research projects on universities, motivated by the student demonstrations at the end of the sixties. Based on these studies, Cohen, March and Olsen developed the notion of the Garbage Can Model.

An organisation is characterised by three general properties: problematic preferences (goals of organisation and participants are inconsistent and ill-defined), unclear technologies (organisation's processes are not understood by members), and fluid participation (time and effort of participants vary).

In sociology decisions are seen as one of the main outcomes of organisations (Luhmann, 1988). The Garbage Can Model discovers, describes and explains failures in organisational decision making processes. It is argued that a decision is the outcome or interpretation of several relatively independent streams within an organisation:

- A stream of problems: Problems are determined by inner and outer organisational circumstances and require attention of participants. Problems are looking for situations in which they might be raised.
- A stream of energy from participants: Participants come and go. It is assumed that they provide energy for organisational decision making.
- A stream of solutions: Members of the organisation produce solutions. Solutions move around, actively looking for questions to which they might be an answer.
- A stream of choices: Choice opportunities represent the point of time when a decision is required by the organisation. Each choice opportunity can be seen as a garbage can into which diverse problems and solutions are dumped.

A special feature of the Garbage Can Model is that not only the participants interact with each other, but also the remaining components of the decision process (problems, choices, solutions) can become active, attract each other, and move away. Thus, this kind of organisation can be viewed as a collection of choices, problems, and solutions. Each component looks for matching other components. According to the Garbage Can Model many different actions are taking place at the same time independently. This provides the model with a high dynamic style.

Now it is time for a few words concerning the striking metaphoric and the main notions Cohen, March and Olsen conceptualised in their model. Firstly the organisation, described as a "collection of choices looking for problems, issues and feelings looking for decision situations in which they might be aired, solutions looking for issues to which they might be the answer, and decision-makers looking for work" (CMO, 1972, p. 2), is called "organized anarchy". Secondly the decision making process takes place in a "garbage can", because one may consider each choice situation as a garbage can into which problems and solutions are dumped by the participants. They do it by chance and with no well-defined intention. Solutions and problems can migrate between the different garbage cans. If a solution meets a choice in the right context and at the

right time, a decision can be made. But the emerging outcomes are diverse and not always as desirable. They can be summarised under three decision styles: (1) If there is at least one problem attached to the choice, the making of a decision leads to a rational outcome (decision by resolution), the problem is solved. (2) Or the making of a decision takes too long and no problems are solved (decision by flight). (3) If the decision is made so quickly that no problem has the chance to come up, it was made by oversight.

The speciality of the Garbage Can Model is not only the comic and pointed name. It deals with the essential sociological task how organisations can survive while struggling with ambiguous and complex problems and an unpredictable environment. The Garbage Can Model turns away from the common view that organisations are the right place for rational, intentional and well-structured decision making in the favour of time and context sensitive behaviour. It is argued that at least parts of any organisation can be described with this model at various times. And in fact, Hickson et al. analysed 150 decisions in British organisations and came to the conclusion, that the form of organisation "is not the primary factor affecting how decisions are made ... More important are the complexity and the policality of the matters under decision" (Hickson et al., 1995, p. 53). Or so to speak, "the matter for decision matters most" (Hickson et al., 1986, p. 248)].

To make a long story short, this is how Masuch and LaPotin (1989) put it: «... reconsider the finale of the James Bond movie 'A view to kill'. Agent 007 balances on the main cable of the Golden Gate Bridge, a woman in distress clinging to his arm, a blimp approaching for rescue. In terms of the Garbage Can Model, the blimp is a solution, Agent 007 a choice opportunity, and the woman a problem. In the picture's happy ending, the hero is finally picked up, together with the woman, and a solution by resolution takes place; the problem is solved. Now imagine numerous blimps, women, and heroes, all arriving out of the blue in random sequence. Heroes take their positions on the main cable. Women cling to heroes, blimps hover above the scene. Heroes may or may not be able to hold an unlimited number of women, but the blimps' carrying capacity is limited; heroes with too many women cannot be rescued. Blimps are retrieving rescuabie, i. e., not-too-heavy, heroes. Women in distress are aware of that and switch heroes opportunistically, choosing the hero closest to retrieval. As women, as well as blimps, make their choices independently of each other, a light hero, on the verge of rescue, may suddenly find himself overburdened. Heavy heroes, in turn, may become rescuable all of a sudden as their women desert them.«

This coming and going is the mechanism called fluid participation. Women may not be saved at all if they change between heroes disadvantageously and all of their heroes of choice turn out to be too heavy; then, these problems are not solved. Heroes may be saved when all women just have left; this is called a decision by flight. Also, heroes can be rescued before any distressed woman was able to hold on to them; then, a decision by oversight has occurred.

Let us come back to the sober grounds of organisational theory and sum up the terminology: the bridge is an organisation, heroes are choices, women are problems, and blimps are solutions. Choices attract problems and solutions. A choice is made if there is an appropriate solution to its problems.[1] Three styles of decision making may appear, but only one of them solves problems.

3 Basic Notions of Reference Nets

Since Petri's thesis (Petri, 1962) many different dialects of Petri nets have been introduced. The basic concepts are concurrency and conflicts, active and passive parts, and the movement of tokens. The few concepts of active (transitions) and passive (places) parts of a Petri Net-system with the restricted relation between them is straightforward and intuitive.

Reference Nets are a high-level Petri net formalism that uses Java as an inscription language. High level-Petri nets are extended by dynamic creation of net instances, references to other net references as tokens, and dynamic transition synchronisation and communication via synchronous channels (Kummer, 1998). They are designed and executed with Renew, the Reference Net Workshop (Renew, 1999), according to Aalst et al. (1999) the only tool supporting the execution of any kinds of nets in nets.

Reference Nets (as Petri nets) consist of three types of elements: places, transitions, and arcs. Semantic inscriptions can be added to each net element. Places can have a place type and arbitrary number of initialisation expressions. On creation of a net instance the initialisation expression is evaluated and leads to the initial marking of the net. Arcs can have arc inscriptions. The arc inscriptions are evaluated when a transition fires and the results leads to the consumption and creation of tokens. Transition may carry diverse inscriptions. There are expression inscriptions which

[1] One might wonder where the participants have gone. In this version participants are not mentioned explicitly. They remain backstage. Now and then they throw solutions into the scene.

are performed when the transitions fires[2]. Guard inscriptions are preconditions to the transitions, i. e. the transition is only activated if all attached guard expressions evaluate to true. Action inscriptions start with the keyword *action* and are only evaluated when the transition fires. Creation inscriptions (consisting of a variable name, a colon, the reserved word *new* and the name of a class net) create new instances of nets.

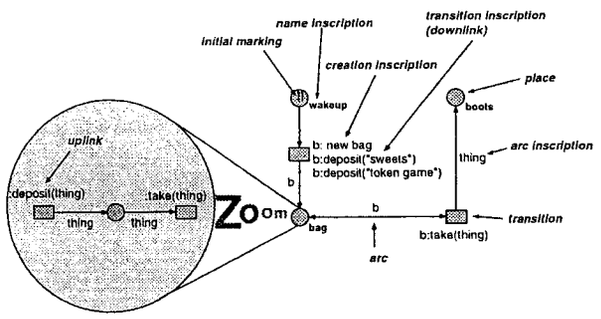

Fig. 1: Sample Reference Net

As known from programming languages function calls can be used for synchronisation and communication. Christensen and Hansen (1992) combined this mechanism with Petri nets by introducing typed communication through synchronous channels for Petri nets. Synchronous channels allow different transitions to be synchronised and exchange data. Both transitions must agree on the name of the channel and on a set of parameters before they can synchronise. This concept is generalised by allowing transitions in different net instances to synchronise. This can only be done, if the initiator of a synchronisation knows the other net instance.

Fig. 1 shows two nets which communicate. The outer net represents the basic schedule of Santa Claus on the night before Christmas. After waking up he takes a new bag and deposits "sweets" and a "token game" into it. Later he can take things out of the bag and put them into children's boots. (Renew, 1999)

The initiating transition must have a special inscription, a so-called *downlink*, specified as a *netexpr:channelname(expr, expr,..)*, which makes a request at a designated subordinate net. The requested transition must have an *uplink* (*:channelname(expr, expr,...)*) as an inscription which serves requests from other net instances. Every time a synchronous channel is invoked, the channel expressions on both sides are evaluated and unified.

[2]Actually the expression inscriptions are evaluated during the search for a binding of the transition. In case the transition does not fire, the result is discarded.

Whenever a simulation is started, new instances of each involved net are created. For any further access on those new net instances now their *references*, which are tokens of other nets, are used.

Reference Nets have successfully been used for system modelling, for agent systems, and business applications, especially workflow systems (e. g. Aalst et al., 1999, Laue et al., 2000, Rölke, 1999).

4 The Garbage Can Reference Net

The Garbage Can Reference Net consists of four net classes: Organisation, Choices, Solutions, and Problems. The *Organisation* (Fig. 2) which is the stage for the elements involved in decision making. It represents the bridge and keeps track of the other net instances and controls the interactions among them. Looking at the *Organisation* the main features of a garbage can decision process become clear: there are the three streams of problems, choices, and solutions pouring into the system. Problems are free until they cling to an available choice. Then switching between different choices is possible. If a solution is obtainable, a decision can be made by removing one choice with an arbitrary number of problems.

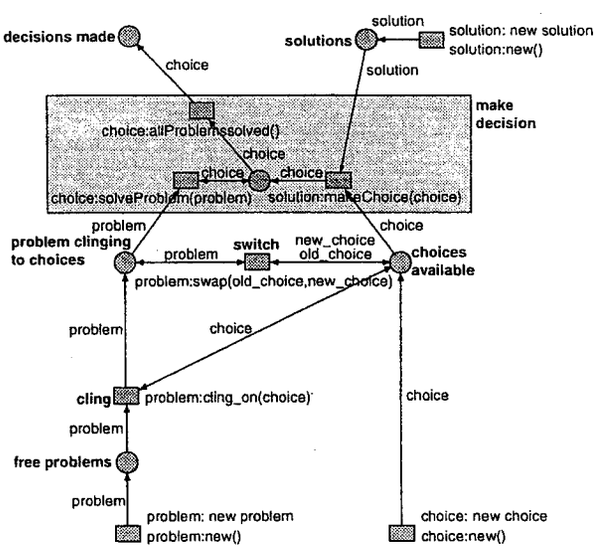

Fig. 2: Reference Net *Organisation*

The *Choices* (the heroes) are the crucial elements of the decision making process and which bring together problems and solutions. The *Solutions* (blimps in the sky, Fig. 3) which bring relief to the distressed situation and lead to decision making. The *Problems* (called women in (Masuch and LaPotin, 1989), Fig. 4) which attach themselves to choices and may be solved eventually. If one takes a look into the net *Problem* (Fig. 5), one can see how a problem can be *free*,

clinging to choice, or *solved* and how states are changed by the transitions *cling_on*, *swap* and *be_solved*.

Concurrency can be found in the transitions *cling*, *switch*, *make decision*, and all the *new*-transitions. They behave totally independently to each other and can switch concurrently to themselves. For sociological theory this means that there is no predefined order in which choices, problems, and solutions appear and interact.

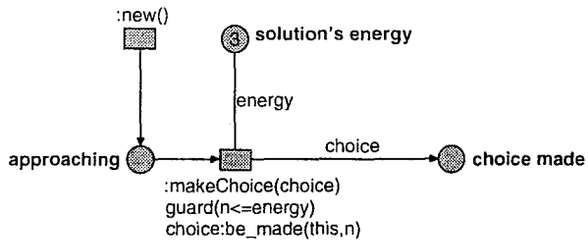

Fig. 3: Reference Net *Solution*

Non-determinism is a key concept of Petri nets. At a given point of time it cannot be determined neither which of the enabled transitions will fire next nor which tokens will be used for the bindings of a transition's variables. In the Petri net formalism for transitions to be enabled it is sufficient that all direct preconditions are satisfied. Thus, information other than local does not need to be considered.

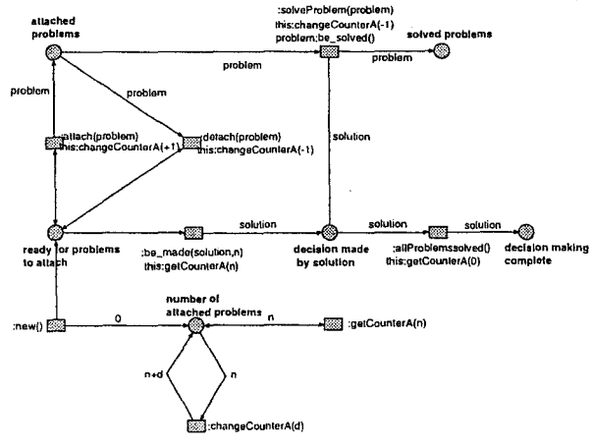

Fig. 4: Reference Net *Choice*

These nets represent a very generalised view on the Garbage Can Model. Apart from the basic behaviour seen here, (CMO, 1972) incorporate aspects of organisational structure, energy distribution among participants and problems, and search strategies for the most attractive choice available. Taking all these features into consideration led to an extended net model with up to 10 net classes (see Heitsch et al., 2000).

Organisational structure controls the access of problems towards choices (which problems may effect which choices) and of participants towards choices (which participants are allowed by the organisation's structure to make which decisions). These regulations give a rudimentary pattern of behaviour to the organisation, but still are far away from total rationality. In a Petri net model these structures limit the amount of possible bindings leading to situations in which a choice can be made, but the available participant is not authorised by the organisation's rules. The problems attached to the choice remain unsolved.

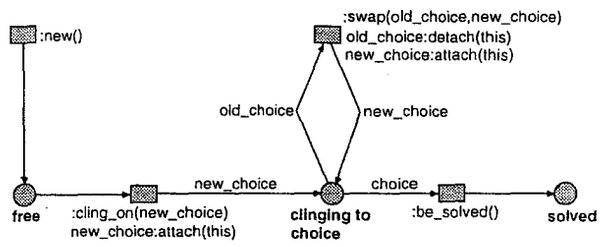

Fig. 5: Reference Net *Problem*

The distribution of energy takes into account the different complexities of problems and the variant skills of participants. On the one hand each problem requires a certain amount of energy to be solved, on the other hand each participant provides energy for problem solving. When the amount of energy available exceeds the energy required, a decision can be made. This aspect is captured technically by changing states of objects which describe if a choice can be made or not.

Search strategies are used to determine which choice problems and participants will select at a given point of time. Problems as well as participants chose the choice closest to decision, i. e. the choice with the least difference of required and available energy. This leads to a »tug of war« between problems and participants. Choices close to decision can either suddenly can be clung to by a large number of problems which prevents decision making or can be processed by too many participants which leads to a decision, but with a waste of energy. Technically, such a search strategy can be implemented by global knowledge of all other objects or by a central instance which acts as a coordinator. (CMO, 1972) require that each problem and participant always is aware of the optimal choice. In the original FORTRAN simulation this was implemented by a simple loop which processes *all* available elements. In

a concurrent system like a Petri net it is more difficult to process all available elements atomically. Tokens move non-deterministically and concurrently through the Petri net (similar to a distributed system). In order to find the optimal choice at each given point of time, a coordinating instance has references to *all* active choices and returns the current »most attractive« choice to the problems and the participants. The global form of knowledge is assumed in (CMO, 1972). Nevertheless, a rather local representation of knowledge as in Petri net semantics seems to be more intuitive for organisations and its members.

In conclusion the Petri net model applies concurrency and non-determinism to the Garbage Can Model and dismisses the necessity of a global clock. Observations of the many different version of the Petri net models result in semantic questions (inspired by terms of Petri nets theory) toward the original sociological theory.

5 Sociological Implications

Cohen, March and Olsen intended to gain new discoveries about decision making processes and their failure in so called "organized anarchies". Their "Garbage Can Model" is laid out as a triad of empirical, theoretical and computational components. Already forty years ago, the authors detected the contribution a computational model can make to the creation of theory. In the words of their research colleagues Kalman C. Cohen and Richard M. Cyert, going back to the year 1961: "The basic advantage of computer models is that they provide a language within which complex dynamic models can be constructed." (Cohen and Cyert 1961, p. 127). But compared with the original simulation of the Garbage Can Model, carried out with FORTRAN by Cohen, March and Olsen (1972) themselves, by all means there are several advantages of the Petri net model. Letting the Petri net model run and taking a look at the outcome allows new insights into for example implicit presumptions of the Garbage Can Model and its link to reality in organisations. In fact, the emerging sociological implications can be classified in four categories, all of them referring to the question of a "good creation of organisation theory".

- The right perspective to organisations, here presented in the form of the sharpened question, how long, how intensive and how extensive an organisation or the respective part of it must be observed?

- The reasonable reference to the reality of the organisational processes taking place and the context of the decision situation. This point deals with the question of taking into account all relevant aspects and not to neglect important issues from the start.

- The formalisation of the theory is the next crucial aspect of theory building. In order to analyse the theory for example with regard to its strong and weak points, it is necessary to decompose the theory in its single components. This is also a prerequisite for the validation of the units in an executable model.

- The possibility to vary the model can lead to new conclusions, relevant to the sociological discourse. For example, the sensibly chosen variation of theoretical assumptions might aim to integrate them in an extended model or mark the boundary of it.

Now let us put these general characteristics in some concrete terms, which emerged from this special approach:

1. One of the main criticism of the Garbage Can Model is addressed to the *inclusion of structures, modes of functioning and patterns of interaction* in the organisation as a whole. Or with the words of Christine Musselin "the organizational context is ignored" (Musselin, 1995, p. 60). Musselin wonders about that matter, because we have not the case "where participants ... have never cooperated with each other before and face a new choice opportunity for the first time" (Musselin, 1995, p. 60). In fact, to isolate the single decision processes from the rest involves some risks. Namely the neglect of "the structure of the relationships between the actors and the possible links between the decisions studied and other decisions" (Musselin, 1995, p. 61). They might then appear not as disorderly as they are regarded now. Furthermore there are no processes like the question of how decision situations and choice opportunities are generated to come into the focus, when dealing with the organisation like Cohen, March and Olsen do. One might get the impression, that Cohen, March and Olsen watched only parts of the 'organized anarchy' and observed the decision making process in the short run. Maybe they decided to do so in order to avoid high complexity. But the reverse of the medal is that many aspects are left out of sight. To sharpen the problematic one might say, that the look Cohen, March and Olsen took to the organisations studied was (1) too short, (2) too partial and (3) too superficial. Dealing with many aspects in this context, which are removing themselves from the directly observation, the modelling with Petri nets comes at the right time. Petri nets can bridge the gap between the theoretical and empirical work in organisational research by providing a tool with

the ability to handle special implicit processes and suggestions and even execute this hidden aspects. This is not a substitution for the proper empirical research but can lead to then definable questions and suggestions and may make the approach to the things happening unintentionally and seemingly unstructured easier.

2. One above mentioned speciality of the Garbage Can Model is the *dynamic aspect* of making a decision, due to the many parallel interactions taking place. Surely there are many actions, which seem to happen unnecessarily and to make no sense. But there is also something like a so-called "power of parallel search" by Cohen (1981): "highly uncertain and equivocal situations can be better explored by boundedly rational agents attacking the problem from multiple perspectives and selecting the best emergent solutions" (Warglien and Masuch, 1995, p. 7). And this phenomena is not an unknown although there are different names given. Lindblom (1959, 1964) chose the name "pluralism" and Thompson (1967) the term "intensive technologies" for almost the same thing. And all of them consider this kind of searching for an solution as a dynamic and creative one. Thus, further research is very promising for the progress in studies of organisational behaviour and can start up with the Petri net formalism supporting the concurrent and non-deterministic aspects. Maybe there will arise something like an "shaped disorder" and an unusual form of "situative social intelligence".

3. The Petri net formalism provides a view to decision processes studied, which is much more true to the theory than the original simulation. For example the fact that new problems and solutions appear by chance, choices are made unpredictable. This view is getting much closer to the implication of the theoretical model, which describes the interactions in a similar way. Beyond, in the original simulation model decisions are always made, when there is enough energy of the participants available. However, in the Petri net model some problems stay unsolved, which one might consider as the consequent pursuance of the principle of non-determinism.

4. There is a big question about processes taking place either in a totally irrational, or in a limited rational, or in a certain rational way according to the Garbage Can Model. Cohen, March and Olsen themselves are dealing with *rationality only implicitly*. Musselin for example interprets the reading of the term of rationality by the Garbage Can researchers as not existing. "To emphasize variations in actors` intentions, suggests that J.G. March et al. concluded that it is pointless to seek rationality in the actions of participants during the decision process" (Musselin, 1995, p. 62). From this statement Musselin derives the criticism that Cohen, March and Olsen deny any kind of rational behaviour from the beginning. So even if there is some, there is no chance to discover it. In Musselin's opinion the term of rationality is understood in a far too narrow way. "Nevertheless, while it is actually harder for participants facing complex situations to elaborate long-term strategies or to anticipate the future, it seems fair to assume rationality in the actor's behaviour, that is, his ability to reformulate the issues at hand in order to have some influence on the process, to seize opportunities or take advantage of the situation whilst it is changing" (Musselin, 1995, p. 63). One can understand and call such behaviour as a *"local" or "situative rationality"*. Petri nets of the Garbage Can model in Heitsch et al. (2000) make the difference between "local" and the common sense of "global" rational behaviour clear respectively discovers an inherent contradiction between the theory and the original computational model itself.

6 Conclusions and Outlook

Computer models provide a bridge between empirical and theoretical work. The requirements of a computer model can provide a theoretical framework for an empirical investigation, and, in return, the empirical information is utilised in developing a flow diagram for the model. Through this process of working back and forth, it is possible to know when enough empirical information had been gathered and whether it is of the proper quality" (Cohen and Cyert, 1961, p. 127). These are again some wise and far-sighted words of Kalman Cohen and Richard Cyert, formulated in 1961.

By applying Petri nets operational semantics were given to the sociological theory. Formal modelling gave explicit meaning to behavioural assumptions which were only made implicitly in the original model by Cohen, March and Olsen. Thus, the formal approach leads to new views on the Garbage Can Model. New ideas of concurrent and non-deterministic behaviour as well as aspects of structure and rationality emerge. If one pursues the goal to deconstruct an existing theoretical model by going into its details and coming out with some new insight to its implications, the Petri net model provides the basis for interdisciplinary discussions, modifications, improvements to the theory, and, lastly, a better understanding how organisations work.

The presented work is one attempt to put a sociological model into a Petri net model. This approach is to be

continued. What we explored by applying Reference Nets to the Garbage Can Model can be extended to other organisation theories and the prevailing views to organisations in common.

Relevant aspects for our view on good socionical theory building are:

- the nature of the view taken on the organisation,
- the dynamic aspects of decision making, which might express themselves as so far unknown "logics of action and interaction",
- the relation between action and structure in organisational decision processes.

The Petri net formalism involves the corresponding flexibility to all these aspects and can help to get this undertaking going. Also Petri nets can be regarded as a relevant and promising tool for the project work coming up. They bring to bear a mode, which is not popular to sociologists and their concepts, even though many of them would like to have it.

As well an organisation as a whole, as a matter of decision, as a group or a single actor can be modelled as a Petri net respectively a Petri net. The impact of organisation theory and connecting ideas to the sociological discourse will be a main challenge we will accept in our future co-operation.

For future research the organisation stands as a relevant miniature of society. Adding the Petri net formalism helps to model, formalise, and verify our theory of organisation sociology.

Acknowledgements

Thanks to O. Kummer, R. Langer, R. von Lüde, M. J. Macdonald, D. Moldt, L. Spresny, R. Valk, and F. Wienberg for the fruitful discussion of our work.

References

W. v. d. Aalst, D. Moldt, R. Valk, F. Wienberg: Enacting Interorganizational Workflows Using Nets in Nets. in: J. Becker, M. z. Mühlen, M. Rosemann (eds.). Proceedings of the 1999 Workflow Management Conference. Working Paper of the Department of Information Systems. 70:117-136. University of Münster, Steinfurter Str. 109, 48149 Münster, Germany, Nov 9, 1999.

S. Christensen, N. D. Hansen. Coloured Petri Nets Extended with Channels for Synchronous Communication. Technical Report DAIMI PB-390, Aarhus University. Aarhus, Denmark. 1992.

K. C. Cohen, R. M. Cyert: Computer Models in Dynamic Economics. In: Quarterly Journal of Economics, 75: 112-127, 1961.

M. D. Cohen, J. G. March, and J. P. Olsen: A Garbage Can Model of Organizational Choice. In: Administrative Science Quarterly. 17:1-25. 1972.

M. D. Cohen: The Power of Parallel Thinking. In: Journal of Economic Behavior and Organization. 2:285-306. 1981.

S. Heitsch, M. Köhler, M. Martens, D. Moldt: Applying High-level Petri-nets to a Theory of Organizational Choice. Technical Report. University of Hamburg. Fachbereich Informatik. Vogt-Kölln-Straße 30, 22527 Hamburg, Germany. in print. March 2000.

D. J. Hickson et al.: Sifting the Garbage Can: Conzeptualizing and explaining processes of strategic decision making. In: M. Warglien and M. Masuch (eds.): The Logic of organizational disorder, pages 35-54, de Gruyter, Berlin/New York, 1995.

D. J. Hickson et al.: Top Decisions - Strategic Decision Making in Organizations. San Francisco, Jossey-Bass, 1986.

O. Kummer: Simulating Synchronous Channels and Net Instances. In: J. Desel, P. Kemper, E. Kindler, A. Oberweis (eds.): 5. Workshop Algorithmen und Werkzeuge für Petrinetze. Forschungsbericht. Fachbereich Informatik, Universität Dortmund. 694:73-78. Oktober 1998.

A. Laue, M. Liedtke, D. Moldt, I. Trickovic: Statecharts as Protocols for Objects. In: Proceedings of The Third Workshop on Rigorous Object-Oriented Methods. York, UK. January 2000.

C.E. Lindblom: The Science of `Muddling Through´, In: Public Administration Review. 19:79-88. 1959.

C.E. Lindblom: The Intelligence of Democracy, New York: Free Press. 1964.

N. Luhmann: Organisation. In: W. Küppers, G. Ortmann (eds.): Mikropolitik: Rationalität, Macht und Spiele in Organisationen. p. 165-186. WDV, Opladen. 1988.

M. Masuch and P. LaPotin: Beyond Garbage Cans: An AI Model of Organizational Choice. In: Administrative Science Quarterly, 34:38-67. 1989.

C. Musselin: Organized anarchies: a reconsideration of research strategies. In: M. Warglien and M. Masuch (eds.): The Logic of organizational disorder, p. 55-72, de Gruyter, Berlin/New York. 1995

C. A. Petri: Kommunikation mit Automaten. Dissertation, Rheinisch-Westfälisches Institut für Intrumentelle Mathematik an der Universität Bonn, Bonn. 1962.

Renew - The Reference Workshop. O. Kummer and F. Wienberg. URL: http://www.renew.de, Release 1.1. October 1999.

H. Rölke: Multi-Agenten-Netze - Modellierung und Implementation eines Multi-Agenten-Systems auf Basis von Referenznetzen. Diplomarbeit. Universität Hamburg. Fachbereich Informatik. Vogt-Kölln-Str. 30. D-22527 Hamburg, 1999.

Sozionik@UHH. Socionics project at University of Hamburg. URL: http://www.informatik.uni-hamburg.de/TGI/forschung/projekte/sozionik, March 2000.

J.D. Thompson: Organizations in Action. Mac Graw Hill, New York. 1967.

R. Valk: Modeling of task-flow system in systems of functional units. Technical Report FBI-HH-B-124/87, University of Hamburg, Vogt-Kölln-Str. 30, 22527 Hamburg, Germany, 1987.

R. Valk: Petri Nets as Token Objects - An Introduction to Elementary Object Nets. In: J. Desel, M. Silva (eds.): Application and Theory of Petri Nets, in: Lecture Notes in Computer Science, 1420:1-25, Springer, Berlin. 1998.

M. Warglien and M. Masuch: The logic of organizational disorder: an introduction. In: M. Warglien and M. Masuch (eds.): The Logic of organizational disorder, p. 1-34, de Gruyter, Berlin/New York. 1995.

Having a Sense of Ourselves: Communications Technology and Personal Identity

Leslie Henrickson
University of California –Los Angeles
Graduate School of Education and Information Studies
Moore Hall, Box 951521, Los Angeles, CA 90095-1521
Lhenrick@ucla.edu

Abstract

As technological advancement accelerates, reactions toward new technologies can elicit resistance and adoption. This paper explores the character of resistance and adoption of technology from the theoretical perspectives of instrumentalism and critical theory. Key to this analysis is the interplay between human senses and technology as it may alter notions of personal identity and of social worldviews. Research in sensory compensation and virtual reality demonstrate both the presumed dominant theory of instrumentalism and the need for a more adequate theoretical grounding, e.g. critical theory of technology. Resistance to technology is examined from the perspective of education. Implications for inter-disciplinary research are discussed.

1 Introduction

> A critical theory of technology requires a substantial vision of what technology is, what it does and what it could do, as well as a normative perspective that provides a philosophical and ethical ground from which to delineate positive and negative forms and uses. (Kellner, 1999)

Moore's Law, a computer industry rule of thumb, reliably predicts that the speed and power of computer performance will double every 18 months.[1] The shape and form of computers has changed as well. Yesterday's room filling isolated VAX mainframe has been replaced by today's desktop internet-connected PC. Tomorrow, wearable computers equipped with the latest sensory devices that send and receive signals from the global positioning system will replace these PC's in our homes and schools.[2]

The rapidity of technological advancement staggers the imagination and catches many people off-guard as they try to absorb the impact of learning new technologies, new tools, new ways of knowing. Reactions toward new technologies can elicit resistance and adoption. This paper explores the character of resistance and adoption of technology from the theoretical perspectives of instrumentalism and critical theory. Key to this analysis is the interplay between human senses and technology as it alters notions of personal identity and of social worldviews. The implications of identity alteration affect both computational modelling researchers and educators. In particular, computational modelling researchers who wish to incorporate socially constructed identity into their models learn that personal identity is not fixed in time or space, and that the use of electronic technology plays a role in such changes. More broadly, for educators the implications are explicitly focused on developing multiple literacies in anticipation of the changing role that human senses play in communications technologies.

2 Instrumental and Critical Theory

A heuristic to guide analysis can be developed as follows. Two types of resistance to technology are an outright rejection of new technologies, and a willingness to use new technologies but under the guidance of the old

[1] In 1958, the first integrated circuit had two transistors and in 1997 the Pentium II processor had 7.5 million transistors.

[2] MIT's wearable computer website http://lcs.www.media.mit.edu/projects/wearables/. The site notes that to date personal computers have not lived up to their name. Wearable computing hopes to shatter this myth of how a computer should be used. A person's computer should be worn, much as eyeglasses or clothing are worn, and interact with the user based on the context of the situation. With heads-up displays, unobtrusive input devices, personal wireless local area networks, and a host of other context sensing and communication tools, the wearable computer can act as an intelligent assistant, whether it be through a Remembrance Agent, augmented reality, or intellectual collectives.

instrumentalist framework. Adoption of technology is also a willingness to use new technologies but guided by a critical theory of technology. The dependency on which guiding theoretical framework is referred to in this dialectic is intimately connected to human sensibility. The resistance-adoption dialectic serves a useful heuristic purpose. Certainly there is a continuum along this dialectic and in some circumstances the same person will be more likely to adopt technology than other circumstances.

2.1 Instrumental theory and identity

Instrumental views of technology are characterized by essentialism, ahistoricism and social abstraction. (Kellner, 1999) The common sense idea is that technologies are tools available to serve the purposes of users. They do not have evaluative content. This means that the use of tools is: indifferent toward the ends of their use, toward politics and ideology; universally applicable in different societies; and, the universality implies that the same set of standards can be used anywhere. Given the instrumental view of technology, the only rational stance toward technology is an unreserved commitment to its use. (Feenberg, 1991)

Characteristic of an instrumental view of technology are conceptions that human identity is something uniquely fixed, pre-given and rationally independent. Individualized instruction, and isolated independent reading and research can best develop a person's identity and proclivities. External guides, e.g. teachers, are not needed in any deeply dependent way but to pass on techniques and practices of know-how. (Stoll, 1995) The affect of tools on personal identity is not about the tool affecting change within us or changing us fundamentally in regards to our perceptual capability. It is only about developing to one's fullest potential what is fundamentally there and pre-existing.

2.2 Critical theory and identity

A critical theory of technology is different from an instrumental view in two important ways. First, a critical theory is historical, contextual, value-laden and concrete. A critical theory of technology recognizes that changes in technology are more than just mechanical improvements to our tools that make our lives more efficient. Technology is deeply embedded in all human interactions, social, work-related, creation of goods and services and culture. Society and technology are in constant flux and, therefore, can never be understood as fixed entities or unique across time and space. Technology is, thus, conceptualized as something socially constructed and imbued with societal biases and interests. Second, a critical theory of technology "follows the dialectical logic of both/and rather than either/or in theorizing new technologies." (Kellner, 1999) It does not set up a false dichotomy between one-sided technophilic or technophobic approaches. Critical theory works with the ambiguities inherent in technology to critique oppressive uses of technology and draw out positive technological implications for enhancing human existence.

Characteristic of a critical theory of technology is notions of personal identity that are socially constructed. Personal identity is in flux and influenced by major societal forces of a historical period. Many refer to the flux as characteristically post-modern. (Kellner, 1995) A post-modern identity refers to multiple identities which one person can assume under different conditions, and refers to new sites and types of identity formation. The affect of tools on personal identity impacts the sites and types of identity formation possible, and does change us in fundamental ways in regards to our perceptual capability.

2.3 Communications technology and identity

The discontinuity between these orientations is more than just a theoretical disagreement between two competing practices. In practical terms, the importance of forming a consistent and coherent theoretical picture of technology, society and personal development cannot be understated. We seek to reconcile our practice and theory in order that we prudently guide our educational practices. If scientific advances outpace an understanding of technology's affect then we will be misdirected in our educational practices, for example, using an instrumental view of technology to guide educational policy. A multi-perspectival inquiry reveals that a critical theory of technology yields a more coherent understanding of technology and society. (Kellner, 1995) My arguments are based on current research on human senses and technology. Consequently, our educational practices must be re-aligned toward multiple literacies as I explain below.

The dominant corporate discourse on technology in education is an instrumentally driven nation-wide commitment to get schools hooked-up i.e. wired and connected, to the information super-highway. The benefits are measured through enhanced learning, empowerment, and increased access to information. Communications technologies are just tools ready at hand to be used at the teacher or students' discretion. Their value is derivative of increased efficiency. (Gates, 1995; Stoll, 1995) Teaching, envisioned thus, can in some cases be reduced to web-page curricular materials that erase the middle-man-teacher.

Arguments that shore up instrumentalism have a disquieting undercurrent. On the one hand, it is the "rugged

individualist" in control of her or his destiny through these tools. On the other hand, communications technology is presented as a redesigned central nervous system connecting everyone in a society to a parasitic relationship with the technology. If such an instrumental vision of the future materializes then this it is no longer just a tool to be used at will, but an essential connection to the living world.

An instrumentalist view reconciles the disjunction between having control of and being controlled by technology calling it a "distortion" in which the individual ultimately wins out. This distortion will go away as we invent more technology that will allow us to control the results we want. Such an unreflective recognition of the impact of technology's affect on human sensibility leads to an infinite regress in justification calling for more and more advanced technologies. For example, distance education video conferencing often creates a spatial distortion that disorients participants. To overcome the spatial distortion and create a more desirable result will rely on advanced research done in sensory compensation and artificial reality. Sensory compensation and virtual reality research, though, provide contrary evidence to this presumed instrumentalist view of technology.

Marshall McLuhan argues that the effects of electric technologies alter our sensibilities in fundamental ways that affect our notions of identity. He uses the terms "closed" and "open" systems, and "inner sense ratio" to describe the phenomena. External tools, or mechanical tools, have extended practically everything a person can do with her body: weapons extend the reach of the arm, glasses extend the reach of the eye, and money is a way of extending and storing labor. Each of these external tools are closed systems within themselves incapable of "collective awareness". On the contrary, our internal private senses are open systems that are "endlessly translated into each other in that experience we call consciousness." (McLuhan, 1995) A ratio of interplay among the private senses, the inner sense ratio, is the response of the body to environmental stimuli. With the speed of electric communications technology, McLuhan argues that we have effectively crossed the border between closed and open systems. Transgressing the border occurs both because of the speed and of the connection to language and consciousness.

One of McLuhan's concerns is the transition that our senses undergo when incorporating new tools into our everyday life. Communications technologies shift the inner balance differently than, say, using a new and improved hammer. The shift in sensibility has an effect on individual identity. McLuhan argues that a change in the inner sense ratio can have aggregate effects on a society. Aggregate effects are reflected through changes in worldviews, conceptions of problems and in social organization. Following McLuhan's logic, the process to understand changes in social structure must start first with understanding the nature of the inner sense ratio, how it changes and what effects obtain on an individual's sense of identity.

3 Cross-talk

Are people talking about the change to our inner sense ratio as a result of new technologies? Taking a multi-perspectival approach to this question, I interrogate sensory science, virtual reality researchers, and a sociologist through their written texts. I find that they do not discuss these topics in a consistent way, nor in a way that presupposes a dialogue or in a way that creates a dialogue.

3.1 Sensory compensation

For decades researchers have dreamed of giving sight to the blind and hearing to the deaf with surgically implanted devices. The blind and deaf themselves, however, have used a different strategy: training another sense to do the job. For example, tactile reading, i.e. Braille, a person can process written information as quickly through the fingertip as someone can by visually reading. Some scientists have followed this lead and devoted their energies not toward fixing the broken "sense" but toward devising electronic devices that help the impaired sense to switch senses more effectively. Lundborg implanted microphones in the nonsensate hand that sent acoustic signals of friction sounds to earphones. Possible identification of different textures was made with acoustics, not sensation or vision. (Lundborg, et al., 1995) Their research demonstrates how malleable the senses can be, how one sense can be substituted to some degree with another.

Other researchers investigate the ability of a sense to transform or retrain itself after being damaged. For example, amputees can feel missing hands grab a cup of coffee, missing feet itch, and missing legs ache. Virtually all amputees experience these "phantom limb" phenomena. In an attempt to explain this, researchers have had to re-evaluate their assumptions about how we engage in the world and learn from experience. That the "phantom limb" phenomena occurs means that the sensory cortex is not hard-wired in but has rewired itself, retrained itself to respond to alternate stimuli. This is referred to as the "remapping theory". Knecht explains that the cortical pathways for the face, hand and torso neighbor one another. Stimulating other areas, e.g. face or torso, can evoke sensations in the missing limb. His research demonstrates, though, that the perceptual changes go beyond what can be explained by shifts in neighbouring cortical representational zones. (Knecht, et al. , 1995) We do not fully understand the open system of our internal senses.

These are two ways in which senses compensate, for one another: replacing one sense with another, and remapping a part of one sense with another part of the same sense. Most of us do not have sense impairments or amputated limbs. So, these events are not common to us and we consider them to be outside of "real" everyday experience. Small groups of people whom most of us would not envy or desire to be in their state. But, is this entirely true that only a small sub-set of people experience sensory compensation? Are there other sets of conditions that allow replacement or remapping to occur? What conditions in the world can simulate the sense replacement and repair?

3.2 Virtual reality

Researchers in virtual reality focus attention on sensory compensation, enhancement and repair, and presumably these effects are only on a temporary basis. Virtual reality is not about providing devices for people to get along in the real world. It's about building imaginary worlds, illusions in cyberspace. Advocates of virtual reality note that to create a believable illusion you only need to provide a few well-chosen cues, "the brain fills in the rest". Virtual reality is beyond the laws of the real world, including gravity, mutual exclusion, distance, size, time.

Japanese researchers in artificial reality concern themselves with the relation of sensory input and output. Research falls roughly into two categories: which sensory cues yield the most comprehensive reality and how to simulate those sensory cues. For example, researchers find a complex relation between sight and sound such that there is a spatial component to both senses. Playing with the spatial component of each sense allows for the simulation or replacement of the other sense. They can induce the same behavioural response by a complex mix of distance, and auditory and visual signals. (Ifukube, 1990) They can induce a change in the inner sense ratio. Similarly, there is work being done on virtual "phantom senses" by studying different elements of tactile stimulus of vibration and temperature.

McLuhan and the two types of sensory scientists have some common ground. All recognize that the inner sense ratio can be controlled, modified or induced by our technologies. The scientific community doing research on the blending of the senses does not find this idea problematic. As a consequence, there is no cultural commentary or critique of the implications of their work, nor is there any mention of the relation of the sensory control to personal identity. McLuhan, on the other hand, calls for this critique of technology's impact on personal identity and its relation to aggregate societal effects.

3.3 Internet

Sherry Turkel is a sociologist of the Internet and has spent the last ten years conducting sociological and psychological assessments of people engaged in heavy Internet use though MUD's or Multi-User Dimensions. Her ethnographic studies focus on the relation between computer use and personal identity formation. There are two things to note about her research. One, she finds the MUD environment to liberate personal identity from a fixed and unitary state into one that allows for "multiple personalities" and fluidity in self-conceptions. (Turkle, 1996) For example, one person can create many different personae on several different MUDs or personae of any shape and form they choose. The self-conception is no longer a core unitary identity but is decentered through the use of this technology.

McLuhan and Turkel would agree that communications technology have an affect on identity formation in these immersion environments. Turkel concludes that the fractured self can emerge in virtue of the technology that is communications based which crosses over the border between external technology and internal sensibilities. Her work provides an insight to the relationship between communication technology and personal identity.

4 Implications for education

The question can be raised what impact does this have for education? There are at least two implications for education. First, we need to develop a deep, fine-grain analysis of resistance and adoption. A first step would be to rewrite the initial heuristic that I introduced. Now, there would be one type of resistance and two types of adoption. Resistance to technology would be comprised of those willing to use technology but under the guidance of an untenable theoretical framework of instrumentalism. Adoption of technology comes in two forms. I argue that an outright rejection of technology is a precursor form of adoption of technology using critical theory, in so far as it is based on a rejection of instrumentalism. It is initially and formally unenlightened about critical theory, but intuitively right on the mark about what needs to be done. The second form of adoption is a willingness to use new technologies but guided by a critical theory of technology.

Current literature focuses on teacher resistance and lack of training as obstacles to technology adoption in schools. This is a very instrumental diagnosis of the symptoms. It may be shown that high resistance to technology integration is based on a fundamental belief that our understanding of technology is not critical enough, that it is too instrumental and untrustworthy. If high resistance is directly related to holding an instrumental view of technology, then to lower resistance en-

tails eliminating an instrumentalist view of technology. Such results would demand that the perspectives change toward regarding technology as socially constructed and embodying historically specific social biases and values. (Kellner, 1998) If this is the case then the initial heuristic of resistance and adoption should be changed as recommended in the paragraph above.

Second, educators must impart the know-how of communication technologies and media literacies that promote the reconstruction of situated knowledges. Education is the lynch pin that provides the tools for people who want to participate in the public and cultural life of the future. Communications tools are essential to all aspects of social life. Multi-perspectival research corroborates that communications technology affects the inner sense ratio and can effect personal identity. Such changes can have aggregate effects on a society. Education cannot turn a blind eye to the technological and scientific advances that are on us now.

Educational practices about new technologies must not only teach the mechanics of how to use the technology but must relay an understanding of the affective nature of the new tools on human sensibilities. This means in part that new understandings of literacy must be developed to meet the challenges of new technologies. (Kellner, 1998) Information content increasingly comes in non-linear forms, e.g. graphical, pictorial and moving images. New forms of content require that new sites and ways of interpreting information be legitimately incorporated into our educational toolkit. As sound, touch and olfactory capabilities, as well as virtual senses, become standard equipment to our communications technologies so too our theories must be accountable to these changed ways of knowing. Moreover, the science behind these standard features will be compensating one sense for another to create the desired alternate reality. Such sensory compensation is occurring without our being aware of it.

We must know about and become aware of how our tools operate on our bodies, the affective nature of what we strap onto our bodies. Literacy no longer can be confined to the linear, alphabetically coded printed page. Navigation will include an intertextual reading between pages, between images, between sounds, i.e. a hyper-"textual" literacy in all these forms. Because of the increased speed and the more direct affect on different human sensibilities that new technologies are moving toward we must develop educational curricular materials and practices that reflect a greater understanding of our tools. To continue on in an instrumental fashion is irresponsible and overtly resistant to reality.

5 Future research

I have noted common research projects within separate disciplines that have bearing on one another, but on which no substantial dialogue has occurred. One reason for this lack of dialogue may be due to the overwhelming instrumental perspective we as a culture hold on the role of technology in society. That is, we generally believe that technology is an inert tool that we have control over. However, electronic communications technologies are fast-paced interactive mediums impinging on our senses with rapidity never experienced before in the history of humankind. The nature of the game has changed.

Sensory stimulation and compensation have direct bearing on our interactions with the world and perceptions of identity. Thus, individual notions of identity can form aggregate societal shifts in worldview. I have reported that no substantial dialogue is taking place that addresses these concerns. I argue that both the lack of dialogue and resistance to technology are due to an inadequate theoretical formulation on the relationship between technology and society. We hurl ourselves headlong into a race with technology as if we were in control because our theory presumes this to be true. The lack of dialogue is evidence for this, i.e. that each discipline can act independently and not have a complex societal effect. The act of resistance is evidence that the theory does not adequately address visceral concerns about technology and personal identity.

Future research can bear out the fine-grained analysis of what has begun here. Future research can begin in a multitude of areas and topics. Some ideas for research in education were mentioned: characterize the nature of resistance, determine skills needed for multiple literacy, and develop curriculum to meet new literacy needs.

Future research to investigate the relationship between our communications tools, and our individual and collective identities can begin both historically and scientifically. Notably, I argue for two criteria across the board. One, that researchers acknowledge their theoretical perspective regarding technology. Two, future research needs to be cross disciplinary in order to create a critical dialogue.

One can review the historical research record in sensory compensation and virtual reality to assess the degree to which an instrumental view guided research and policy formation. The historical record could be recast in light of the contrast between an instrumental and a critical perspective toward technology and society. Observed trends over time may inform future research.

Cross-disciplinary discussion should take place between researchers in sensory compensation, virtual reality and communications. Other researchers would also be interested in this topic, e.g. psychologists. Reframe research questions aligned with a critical theory of technology. In general, make theoretical assumptions clear. In particular, some questions to ask are: To what extent do particular technologies effect personal identity or notions of who we think we are? How will we measure this?

It's likely that computational modelling will play an increased role in policy decisions for complex social problems. This is due to advances in both computational theory in the forms of chaos and complexity theories, and advances in high-speed computers that open up new realms of quantitative exploration. These new theories and computational techniques lend themselves to social science inquiry, the inquiry into relationships, networks and decision processes of humans with identities that undergo change.

First and foremost, what is at stake is how we theorize technology and, second, how to evaluate the costs and benefits of technology for society. Understanding the development of personal identity as both socially constructed and as informed by the electronic tools we use will be important factors for incorporating the social dimension in computational models and future research. Both educators and researchers can play a pivotal role as critical guides about new communications technologies and in how we will come to know ourselves in relation to the tools we use. To the extent that educators and researchers can do this is largely based on their theoretical perspective of technology and society.

Acknowledgements

I am grateful to Douglas Kellner for his insightful suggestions and comments.

References

Feenberg, Andrew. *Critical Theory of Technology*. Oxford University Press, New York, 1991.

Gates, Bill. *The Road Ahead*. Penguin Books, New York, 1995.

Ifukube, Tohru. Proposal of a New Method to Evaluate Vertigo Based on Induced Motion of Vision. *Japanese Journal of Medical Electronics and Biological Engineering*, 29(4):15-20, 1990.

Kellner, Douglas. *Media Culture: Cultural Studies, Identity and Politics Between the Modern and the Postmodern*. Routledge, London, 1995.

Kellner, Douglas. Multiple Literacies and Critical Pedagogy in a Multicultural Society. *Educational Theory* 48(1):103-22, 1998.

Knecht S. and H. Henningsen H, T. Elbert T, H. Flor, C. Hohling, C. Pantev C, E. Taub. Reorganizational and Perceptional Changes after Amputation. *Brain*. 119:1213-1219, 1996.

Lundborg, G. B. Rosen, B. and S. Lindberg. Hearing as Substitution for Sensation: a New Principle Forartificial Sensibility. *Journal of Hand Surgery-American*. 24A(2):219-224, 1999.

McLuhan, Marshall. *Essential McLuhan*. Harper Collins, New York, 1995.

Stoll, Clifford. *Silicon Snake Oil*. Anchor Books, New York, 1995.

Turkle, Sherry. *Life on the Screen: Identity in the Age of the Internet*. Touchstone Books, New York, 1995.

Modelling Agent Systems Using the Hotel Analogy

"Sanitised for your Protection"

Lindsay Marshall and Savas Parastatidis
Department of Computing Science,
University of Newcastle upon Tyne, Newcastle NE1 7RU, UK
Lindsay.Marshall@newcastle.ac.uk, Savas.Parastatidis@acm.org

Abstract

This paper looks at how a particular social analogy (that of the hotel) could be used to help the design of the environment provided by an agent support system. It discusses some of the implementation issues and problems that the use of the analogy exposes

1. Introduction

Analogy is a wonderful tool for finding new models and approaches in computing, and it is useful at all levels from the most conceptual down to the practical. Social analogies in particular are revealing when exploring various aspects of the design space of agent systems. Much insight on how agents could and should interact can be gained from this. However, our concern in this paper is not with agents and their interaction, though we will have something to say about this. For more information on agents the reader is referred to (Genesereth and Ketchpel 1994; Franklin and Graesser 1996; Maes 1994; Maes 1995). We are principally concerned with the environment that supports the agents and how the agents interact with that environment. An environment that must be dependable (i.e. secure, reliable, available etc.), both from the point of view of the agents' owners and that of the service providers on whose computers the agent programs execute.

An early experiment lead to the development of the *Iris agent execution environment* (Parastatidis 1996) which provided an *office*-like model for the execution of agents. Software components such as *secretaries*, *receptionists*, *managers*, *security advisors*, *messengers* constituted the building blocks, or the *personnel*, for *branches* and *agencies*. The collection of all of the branch and agency offices formed the Iris agent execution environment. The two different types of offices provided distinct services to the visiting agents. The organisation of the personnel in each office resembled the organisation of a human office system where each member of staff has a specified range of duties.

Evaluation of the Iris agent execution environment lead us to consider other support system analogies and having looked at various options (e.g., libraries, schools, public transport) we decided that one of the most potent is that of the *hotel*. It is this analogy that we will explore in the rest of this paper. Anyone who has ever stayed in a hotel should grasp the reasoning behind our suggestions without difficulty, and be able to see many other extensions that we have missed or which we do not have the space to cover. We see the hotel analogy as a good design model for future agent-based software architecture systems and we show how it can highlight potential areas of difficulty that implementors must consider when working. Please note that much of what we say may seem blindingly obvious. We would (in most cases) maintain that if something is obvious, it is because of the power of the analogy.

The rest of this paper is organised as follows. Section 2 introduces the notion of the hotel while Sections 3 to 8 discuss the design analogies with common actions in hotels. Additionally, we look at potential problems with human guests in hotels that have analogies in the agent world. Finally, Section 9 presents our conclusions.

2. The Hotel

Before we proceed in describing the hotel analogy for an agent system, we first need to define what an agent is. For the purposes of this paper, we define an agent as *a program that relocates itself from host to host, carrying out computations at each place it visits*. The exact nature of these computations is not important, but clearly it does provide the reason why particular hosts are

chosen. We will see later how the choice of particular hosts may be influenced.

In our analogy, we regard each host as a hotel in which the agent acquires a *room* for a specified, but possibly extensible, length of time. The room provides the controlled environment in which the agent can execute its tasks. The environment consists of a predefined collection of resources that the agent requests during the check-in process (Section 3). Agents may call for additional resources through the hotel's *room service*. Of course, additional resources are charged to the agent's room. Room service becomes the medium for the interaction between agents and hotels (more on room services in Section 4).

The behaviour of an agent resembles that of a traveller. The agent moves from one hotel to the other acquiring rooms and consuming the resources provided. Also like a traveller, an agent may bring with it *baggage* that provides additional materials that the agent needs to work, but that are not provided by the hotel – when staying in a hotel you usually take your toothbrush, but you will probably not take towels as they will be provided. This baggage is the private property of the agent and as such much be maintained securely and safely.

3. Check-in

When an agent arrives at a host where it wishes to operate it must first check-in. (Clearly the check-in process could be null and thus our hotel simply the equivalent of a crash pad, but this is definitely a degenerate case and we shall not consider it further). The queue of agents is processed in order of arrival unless the agent holds VIP status and therefore is subject to special treatment. When the agent's turn arrives it must identify itself and arrange for payment. As always the issue of identity is interesting. Some hotels may operate like a members only club and in these establishing the identity of an agent, and therefore its status as a member or member's guest, is vital. But in other places, it turns out that identity is less important than it might at first seem, as payment is the main issue – after all, you can check-in to most hotels using a false identity as long as you can pay! And then of course there are the sorts of hotels where they are most definitely not interested in your identity at all...

The simplest case is where the agent has a pre-booked room. Having identified itself and thus been associated with the booking, payment must be arranged. It may be that the booking includes billing information already: "Charge company X." However, the agent might have to proffer some kind of e-cash or other credit token which the host will attempt to authenticate. A suitably intelligent agent may also wish to haggle in order to negotiate a better price ("I'm a member of the AAA1"). The booking will have pre-specified the type of *room* required by the agent, but at this time another type or additional services can be requested ("Can I have a Guardian in the morning please?"). A host can provide support for different kinds of agent through *themed* rooms, thus it could have Java rooms or tcl rooms depending on the implementation base of the agent.

Note that so far this whole process is similar to what went on in batch processing systems where jobs were submitted on cards to the system; JCL cards carried information about the resources needed by the program that followed and the operating system used this information to schedule and control the program.

If no booking has been made, the agent may be turned away because of a lack of rooms or because the hotel will only deal with pre-booked agents. On the other hand, if the agent appears able to make payment for the length of stay requested then the hotel may allow it to remain.

Once the hotel is satisfied that any resources used by the agent can be paid for, and that a suitable room is available, a room will be allocated and the key given to the agent. The room is vital as it provides the agent with the support facilities that it needs to carry out its operations. A hotel may provide many kinds of rooms with prices to match, or, like a Tokyo capsule hotel, one kind of room at a flat rate. It all depends on the service that hosts wish to offer. Clients that pay more will (usually) get better facilities. For instance, a host may provide a stock information service to rooms and will delay providing this information to the occupant by an amount dependant on how much was paid.

If the agent has visited the hotel before then the hotel may know of any special requirements it has and provide them automatically. A hotel may, for example, maintain in its customer database special requests for room services, preferred types of rooms, usual method of payment, etc. The hotel may also keep track of its regular customers and provide them with discounted rates and extra services. Hotels may be parts of *chains* and information about regular users may be centralised and available to all members of the chain.

4. Room Service

As indicated above, the level of *room service* an agent gets depends on the amount paid. But what kinds of services might be available? Agents do

not have the same needs as we do! The most obvious service is on a par with a hotel room having a bed—access to CPU power. The agent is there to compute and the amount of CPU time it gets in a given elapsed time will relate to the kind of room it gets allocated. There is no reason why agents should travel alone—you can have the equivalent of double and family rooms for multi-process agents.

Most hotel rooms provide access to information and ours are no different. There will be some free services (for instance date and time), some standard chargeable services (sending and receiving messages over the network) and the equivalent of *pay per view* movies or the *mini-bar* where special access to information is given and a premium charge levied. (Such as the stock market information mentioned above) Once again the level of access and charge rate will depend on the kind of room requested.

Agents can expect any servicing of the room to be invisible and that all their transactions with the system and the outside world are secure, confidential and private. The host should provide facilities whereby an agent can be notified of and receive incoming messages which are addressed to it care of the hotel – note that some of these may come before the agent has arrived or after it has left, these cases must be handled properly with storage, forwarding or just by simple return.

5. The Stay

During its stay at a hotel, an agent may only use the resources available in its room. It is not allowed to access any of the resources in other rooms (the privacy of other agents would be compromised otherwise). However, an agent may *invite* other agents to its room. The invited agents may provide additional functionality or just use the room's available resources to communicate and exchange information with each other. In any case, the hotel is not responsible for the invited agents. The hosting agent will have to pay for any additional resources required and also manage the available resources and communication.

A host could also provide for *conference*-style meetings of agents. There may be situations where many agents need to participate in a common task or make information available to each other. For an additional fee the hotel may provide a special room with its own facilities where the agents can meet to work together.

Real travellers rarely stay in their hotel all the time, unless it is a resort hotel or a conference centre, and it is not unreasonable to think about an agent using a host as a local base for information collecting which it then either processes in its room or takes with it in its baggage on leaving. There may be sound reasons, for instance bandwidth or security considerations, why information may be best accessed using the hotel as a base rather than doing it more remotely.

When dealing with truly social agents that independently interact with others, hotels almost always have bars and restaurants where people meet, and our environment could provide similar facilities for agents where they could interact outside the confines of their room. Quite what this implies depends entirely on the nature of the interactions that the agents themselves can support.

6. Support Issues

Hotels are never without troubles from their residents. We expect that malicious or badly behaved agents will exist in our hotel-based agent execution environment and they must be dealt with appropriately. Equally guests often have trouble with their hotel and mechanisms for dealing with this must also be in place. Careful consideration of the workings of a real hotel can reveal many areas where things can go wrong in the hotel/guest interface and the guest/guest interface and we have identified a few of them in this section.

6.1 The 007 Problem

Agents may attempt to use their rooms as the basis for spying on other agents in the same hotel. They may try to monitor their activities, the information they hold, the resources they are using. Hotels should not allow this to happen – as we noted above all transactions should be secure, private and confidential. This means that the system must be designed to eliminate as far as possible loopholes and covert channels – there should be no electronic equivalent of bribing the maid to get a key to another room, there should be no connecting doors unless explicitly requested.

6.2 Rock Star Problem

We expect that there will be agents that will try to trash the rooms they book and maybe the whole hotel. This will manifest itself in the form of excess resource consumption and means that scheduling and monitoring systems must be carefully designed to prevent this happening, or else to ensure that the host charges the culprits appropriately for what they consume and that other affected guests get some kind of compensation. This kind of problem suggests that an interesting area for exploration is the provision of insurance to services that support agents. After

all, real hotels are all insured against a wide variety of risks so why not a virtual hotel? We are not aware of any company that currently provides this kind of protective insurance to companies.

6.3 Barton Fink Problem

What happens if the agent in the room next door turns out not to be as nice as they seem? What does an agent do if it finds something not quite right about its environment? Chains of responsibility must be clearly laid out so that such situations can be investigated and any problems contained. In a hotel environment, there is a web of levels of trust that covers every part of the system. Guests assume that the hotel staff are not stealing from their rooms or looking at their private material. They are probably less trusting of their fellow guests, though they may become more so if they meet. The hotel probably trusts no one.

6.4 Garbologist Problem

Another approach to maliciously acquiring information from agents is to examine their execution environment after they have finished. There may be malicious agents that manage to book the same room in a hotel as a target agent that has just checked-out. By examining the room the previous agents left they may manage to acquire important information (e.g., examining the registers of a CPU or reading the memory associated with a particular room). The hotel must activate a *cleaning service* immediately after an agent checks-out that removes any *garbage* an agent leaves behind. Of course the agent may not trust the hotel to carry out this task effectively enough and may itself try to ensure that no traces remain.

6.5 Hotel California Problem

What happens if an agent checks out but cannot leave? That is, the agent's ability to move around is compromised by the host. In this state the agent will appear to its owner as having vanished as it cannot communicate, after all it has checked out so has no access to host facilities. If the host is malicious it can answer queries about the agent by saying that it has moved on even though it has not. Some agent programs may be valuable and thus be subject to theft or ransom (codenapping).

7. Check-out

When an agent leaves a hotel it should check out. This provides a point where actual resource usage over and above that agreed at check-in can be identified and charges assessed. The agent must negotiate payment for these extras. Equally, it allows the agent to verify that it (or its initiator) is being charged only for those services that it has used and no others. If it appears that resources that should have been provided have not been then a refund should be asked for at this point.

What happens if an agent has used services that it cannot afford? In a real world, commercial system this could be a real problem – after all a program cannot wash the dishes to pay for a meal! Once again insurance would help, but we would foresee that this would be a stimulus to the increasing provision of member only services where, as we indicated above, the identity of an agent, and thus of its originator, is clearly established at check-in. Knowing this may (though not always) allow recourse to conventional legal means of redress.

Once checked out the agent move to another hotel to carry out more work and if the current hotel is part of a *chain*, it may be possible to have forward bookings arranged by the receptionist instead of the agent having to do this itself. This may potentially give it more privileged treatment at the other hotel – the regular customer analogy.

8. Choosing Hotels

Based on the services and the resources the hotels provide, agents can choose which one they want to visit. Thus there will be five star hotels which are expensive but provide the equivalent of unlimited luxury – lots of storage and CPU power, or perhaps some special purpose hardware (though seems more the province of theme hotels). Equally there will be the equivalent of commercial hotels and theatrical boarding houses where agents that travel regular routes can stay economically. The guarantees about security and privacy provided by the hosts will also form an important part of any decision as to which to use.

Clearly the existence of chains where you can expect a known level of service at each member host will be useful when planning the route an agent needs to take to achieve its ends. The chains themselves need not be monolithic organisations, they could be fairly loose groupings of hosts that conform to an agreed set of standards – such collections of independent hotels exist in the real world, though they often tend to be in the more expensive sector of the market.

Ultimately, the choice of where a operator decides to send an agent will depend on many factors, most important being the function that the agent has to perform – there may be only one place that provides the required facilities! A truly independent agent would have to written so that it could evaluate which hotel would be best for it, a challenge that many humans find hard and perhaps another good reason for creating chains of hosts.

9. Conclusions

When designing and implementing a new system it is always useful to have an existing model which can provide insight into the kind of problems that one might encounter. Clearly it is essential that an appropriate model is chosen, and one that is sufficiently rich that it will allow the consideration of a wide variety of possibilities. As we stated above, we believe that the hotel analogy provides just such a model for developers of agent systems. Our brief tour of how guests use hotels, how hotels service guests and how this could be applied to agent systems shows, we believe, that the analogy is as powerful as we claimed it to be. We have only really touched on each topic and there is considerable room for the expansion and development of the idea. Everyone's experience of hotels is different and will illuminate different aspects of what is needed from a support system.

Acknowledgements

Our thanks go to Professor Santosh Shrivastava who first mooted the idea of using the hotel analogy.

References

Franklin, S. and Graesser, A. (1996). "Is it an Agent, or just a Program? A Taxonomy for Autonomous Agents." In *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages.* Springer-Verlag.

Genesereth, M. and Ketchpel, S. (1994). "Software Agents." *Communications of the ACM*, 37(7): pp.48-53.

Maes, P. (1994). "Modelling Adaptive Autonomous Agents." *Artificial Life Journal*, 1(1 & 2).

Maes, P. (1995). Designing Autonomous Agents. MIT Press, Cambridge.

Parastatidis, S. (1996). *The IRIS agent execution environment.* MSc. Thesis, Dep. of Computing Science, University of Newcastle upon Tyne.

The Making of Meaning in Societies: Semiotic & Information-Theoretic Background to the Evolution of Communication

Chrystopher L. Nehaniv
Interactive Systems Engineering
University of Hertfordshire
College Lane
Hatfield Herts AL10 9AB
United Kingdom
C.L.Nehaniv@herts.ac.uk

Abstract

We examine the notions of meaning and information for animals or agents engaged in interaction games. Concepts from cognitive ethology, linguistics, semiotics, and evolution are surveyed. Innateness, individual learning, and social aspects (social learning and cultural transmission) of the evolution of communication are treated. Studies on animals and agents showing degrees of communication are analyzed with an eye to describing what aspects of communication actually are demonstrated, or also in the case of many simulation studies, are built-in to the system at the outset. In particular, predication and constituent structure (subcategorization) have so far never been shown to emerge in robotic or software systems.

1 Introduction

Meaning in real human societies is socially constructed (Bruner (1991)), yet this depends also an the individual member of society's participation. The making of meaning in society emerges from the interaction of many participants as they communicate to one another about the world in which they are situated. Obviously the participants have particular biological capacities necessary for a construction of meaning, but the degree to which innate mechanisms as opposed to learning or cultural mechanisms are involved is the subject of much debate, especially in the case of human language acquisition. For other animals and for software and robotic agents, evolving or designing communication systems present similar issues. The substrate upon which communication relies can be compared and contrasted to the human case, and the insights should be useful in several areas: (1) understanding human communication and language situated in the context of a general biological background, (2) identification and description of characteristics, properties, and mechanisms sufficient for the support of communication systems of various kinds in animals, (3) the design and construction of mechanisms to support communication and language-like phenomena in artificial systems.

Semiotics provides an insightful approach to understanding meaning in terms of a *relational* (rather than a naive *mapping*) framework (Peirce (1995); Goguen (1999)). In particular, a *sign* or signal is related to a *signified* via an *interpretant*, the situated linkage between the two, depending on participant in the particular act of *semiosis*. The segregration of the *sign* and also of the *signified* from the background of the environment are not a priori given, nor need they coincide for different participants in an act of semiosis. (Although this theory of meaning seems simple enough, it is much more complex than what one usually sees in agent studies of the "evolution of language" or "evolution of communication", which assume a (generally fixed) set of possible referents and (generally fixed) set or alphabet of signs, both available to all agents at the outset.) The legs of the semiotic triad (sign, signified, interpretant) all vary with the particular agent in question. Thus the study of meaning is inherently an agent-oriented research area, rather than a third-person God's eyeview Platonic world of absolutes.

An information-theoretic approach can be used to study the evolution of channels of meaning in a community of agents (Nehaniv (1999)). At a fundamental level, modes of sensing and actuating afford an agent its access to potentially meaningful information – *meaningful information for a particular agent* is information that is, in a statistical sense, *useful*

for the agent in achieving its goals. In addition to the interaction channels, internal structure and history of the agent also plays a critical role in facilitating the use of meaningful information to achieve its goals. Applying Shannon information theory (Shannon and Weaver (1963)) to information in channels that are meaningful in this sense allows one to develop an agent-based theory of meaning as an extension of the mathematical theory of communication.

The foregoing remarks already apply to a single agent or animal interacting with its *Umwelt*, the ethologist's term for the environment in which it is embodied and embedded. Moreover, for social animals, and for socially intelligent agents, meaning (in the sense just outlined) emerges from the interaction of semiotically active agents comprising the society. Which goals are desirable for an agent depends on its nature, but also the culture in which it developed, channels of information that are meaningful for attaining these goals are in part determined by design (evolutionary or intentional) and in part by the history of interaction with others. The segregation of signs and signified from a morass of environmental stimuli to comprise legs of the semiotic triad (within components of a system of signs) depends also on embodiment, society and history of interaction.

We argue that useful models of the evolution of communication must take into account the principles described here, and that other current models of fatally flawed methodologically or at best incomplete. Indeed many published results in the evolution of communication can be shown to be consequences of random statistical sampling errors leading to convergence of (naive) "communication systems" in which the potential signs and their referents were circumscribed by experimenters at the outset and in which a (naive) notion of semantics constrained the nature of the possible systems which could evolve — only in a manner that would seem to confirm the preconceptions of the experimenters. Similar remarks apply to the "emergence of syntax" in which constituent structure (essentially context-free language formalism) has been built-in at the outset (e.g. as "slots" in semantic processing).

How socially and semiotically realistic study of the evolution of meaning could be carried out will be described, making reference to some fundamental studies in the ground of communication (Wittgenstein (1958, 1968); Billard and Dautenhahn (1999); Dautenhahn (1995); Nehaniv et al. (1999)).

We will throw out several assumptions that are in made with traditional denotational semantics, by making contrasting assertions:

1. **No Agent, No Meaning** It will not be possible to have a God's eye view notion of meaning. A signal or message can only be meaningful for particular individuals involved in particular interactions with their environment or with each other.

2. **No Privileged Meanings** We will not assume there is a special set of concepts and predicates that characterizing the set of what it is possible for the agent to mean. This is for instance a rejection of a Platonic realm of forms, which the real world is only a shadow of, etc. Moroever, this entails that we may not assume a priori that certain categories (classes of objects, attributes, abstractions, etc.) exist outside agents and their interaction. The existence of such categories must always be grounded in the particular agent's internal architecture, e.g. the state of its neurons, etc., as they relate to its previous experience and interaction with others and the world.

3. **No Privileged Signals** We will not assume that there are specific, atomic symbols or classes of symbols to which all agents may in principle have access. The sensory and actuator characteristics, as well as learning and experience, conspire to determine what type of event constitutes a signal for the particular agent in question.

4. **No Privileged Mapping** Agents may have incomplete knowledge of symbols and referents, actions, meanings, that might be communicated. Moreover no particular mapping of signals to signifieds is the privileged correct mapping. Agents may have different and conflicting mappings, with different domains and ranges.

Thus each vertex of the semiotic triangle is subject to variation. Different agents use different interpretants (hence potentially different mappings) to relate sign and signified.

2 Semiosis: The Making of Meaning

A much less naïve theory of how meaning arises than the denotational semantics common in computer science is semiotics (Peirce (1995)), introduced by an American philospher working over a hundred years ago.

2.1 Semiotic Triangles

Semiotics acknowledges the situated nature of the making of meaning. The connection between a *sign* and what it signifies (the *signified*) is mediated by an *interpretant* (the relation between them). The naturalness of this relationship has degrees: A sign may

be iconic (sensorially indicative of the signified), indexical (indicative but not representing the signified in a way closely matching the perceptual stimulus the signified would produce), or symbolic (arbitrarily associated to the signified). Examples of iconic signs include threatening displays in animals, indexical signs include the intention movements of animals or a hole in a wall indicating that a bullet passed through it, and symbols include arbitrary phonemic strings of spoken human language.

By making the interpretant explicit, Peirce made clear that the relationship between sign and signified is not a static one, it can vary with the agent involved, and between agents. Sign, signified, and interpretant are vertices of a triangle on which each process of making meaning is based. Such a process is called *semiosis.*

The above rejection of the assumptions of denotational semantics and similar systems amounts to recognizing that each aspect of semiosis — sign, signified, and interpretant — is thus agent-particular rather than part of some external structure.

3 Meaning is (Statistically) Useful Information in Channels of Sensing and Actuating

We now relate the semiotic notion of meaning to its situated and embodied contexts in human, animal, and other agent systems.

3.1 Wittgenstein and Meaning in Use

Denotation of words may be relatively unambiguous for proper names, but general concrete terms, actions, attributes, and relationships correspond to no particular entities in the physical world.

Wittgenstein pointed out that to know the meaning of a word one must know the function of the word in the contexts in which it is used. Generalizing from his insights, we shall insist that the meaning of signals can be and should only be defined in terms of their usage in *interaction games* (Nehaniv (1999)). Animals do not evolve signal systems for the purpose of making 'true' assertions about the physical world. They are not concerned with truth, but rather with survival in the natural world. If they can use signals to manipulate the world and gain useful information about it, then this is meaningful for them and can motivate natural selective pressure.

Meaning is understood here as (1) information in interaction games between an agent and its environment or between agents mediated by the environment and in all cases by the sensors and actuators of the agents, and as (2) *useful* (in a probabilistic sense taking into account the costs and benefits of sensing and actuating) for satisfying homeostatic or other drives, needs, goals, or intentions. (see also Nehaniv (1999), Nehaniv et al. (1999)).

3.2 Private Meaning

The definition of meaning above is made with reference to a particular agent (or possibly a community), since the notion of "useful" requires this and since the notion depends also on the particular sensing and actuating capacity of the agent. Thus information that is meaning for one agent may be imperceptible or meaningless noise to another. Moreover, the internal state and structure of the agent is crucial to whether information might be useful to it. This is closely connected with whether the agent can use the information to modify its *expectations* (e.g. predictive scenarios) of what is likely to happen and thus modify its own future actions in light of these. (Also compare the discussion of Smith (1996) below).

4 Evolution of Communication

Darwin (1872) recognized the importance of the expression of emotion in an animal as cues by which others can judge aspects of its internal state, and thus its likely future behaviour. Cues, communicative signalling, and misinformation are distinguished in the literature on animal communication and information-theoretic properties are related via cost-benefit tradeoffs to the study of the evolution of communication.

4.1 Definitions of Communication

(Bradbury and Vehrencamp (1998)) define communication as follows: "The process of communication involves two individuals, a sender and a receiver. The sender produces a signal which conveys information. The signal is transmitted through the environment and is detected by the receiver. The receiver uses the information to help make a decision about how it should respond. The receiver's response affects the fitness of the sender as well as its own. In true communication, both sender and receiver benefit (on average) from the information exchange."

Stimuli produced by an animal but not benefitting it perceived by others are called *cues.* If the production of the signal does not on average benefit the receiver, then this is called *misinformation.* Examples include the mimicry of one species' sexual pheromones by another in order to attract the former as prey, the use of fishing bait, but also camouflage and disruptive displays in animals (e.g. cephalopods Moynihan (1985); Hanlon and Messenger (1996)). (Misinformation is sometimes called "dishonest communication", but we avoid this term in that it leads to

presuppositions that the receiver is capable of holding a false belief or that the emitter intends the receiver to form a false belief, etc.) Signals may be very extended in temporal extent, *states* (e.g. permanent coloration markings on the body, fixed body scents) or *events* of more limited scope (alarm calls, a display of out-spread tail feathers, aggressive posturing and coloration, etc.).

Many definitions, not requiring benefit on average to the recipient, of a signal occur in ethology:

"Communication is the phenomenon of one organism producing a signal that, when responded to by another organism, confers some advantage (or the statistical probability of it) to the signaler or his group." (Burghardt (1977))

This definition is used by MacLennan (1992) in a synthetic computational ethological implementation. Populations of "simorgs" (essentially look-up tables giving functions from global environment and local environment to either emissions and actions) are subjected to digital evolution in which they are rewarded for actions matching the local environment of the last emitter. Comparing evolution (using a steady-state genetic algorithm) of such simorgs to others for which communication was not permitted, MacLennan showed that Burghardt's definition is satisfied.

4.2 Expectation, Prediction, and Action

(Smith (1977, 1996)) considers that an animal's basic cognitive activity is characterized by "a continuous cycle of generating and testing expectations that are incorporated into predicative scenarios". Expanding this: The animal is seeking or extracting information from various sources, in various circumstances; it compares this information with information it has previously stored; and it makes and updates predictions, selects among them and generates new ones. This is a continuous process, in which information is used to produce expectations. Signals from other animals is an important source of such information. The information and predications of an individual are largely "private", i.e. not visible to others, but may be made public by specialized behaviour called signaling, e.g. information about what the individual is likely to do next. Signaling behaviour can influence the recipient's behaviour in a manner that is useful to the sender. The behaviour of populations that signal will co-evolve with the dispositions of how recipients respond whether the recipients be in the sender's own population or another allospecific group. Formalization of signal repertoires, specialization of displays, modes of varying display form, modes of combining displays, and formalization of interactions will all be driven by the costs and benefits of signaling behaviour, and are especially likely to have effects on recipient expectation of social events (Smith (1996)). Moreover, Smith emphasizes that formalization of signalling interaction enables each participant to elicit signalling responses within formal (and thus more predictable) constraints. Here we have the evolution of interaction games via the formalization of signalling exchanges.

The communication behaviour here arises in an evolving populations engaged in social and nonsocial interaction. The nonsocial components have to do with manipulation of the environment, of predator, and of prey; while the social component can be largely (but perhaps not completely) identified with intraspecific interaction (territoriality, mate attraction, etc.). Cues such as direction of eye gaze and joint attention or signals of intention movements may be interpretable across several species, and might be considered candidates for interspecies communication (subject to further conditions of the various competing definitions).

4.3 Communicative Systems

Animal communication thus is clearly subject to inherited genetic and developmental factors. Innate signalling systems might be refined by experience, e.g. young Vervet monkeys may make inappropriate alarm calls, ignored by adults, before they can distinguish harmless birds from aerial predators, (Seyferth and Cheney (1986)). Chomsky (1968, 1975), Pinker and Bloom (1990), Bickerton (1990), and (Maynard Smith and Szathmáry, 1995, Ch. 17) have argued that human ability to acquire language is biologically based or innate. In particular features of the ambient language's grammar are acquired by setting parameters in a universal grammatical system for human language (Chomsky (1981)). This system might be inborn or developed, in that all humans acquire it in the course of development, and may have a large genetically transmitted component that is not merely part of general cognitive abilities and intelligence. Meanwhile, others argue that general human cognitive abilities will eventually be able to explain the origin and maintenance of language (e.g. Steels (1995)). Many workers are now studying the degrees to which innateness or competing mechanisms can serve as a explanations of the evolution of linguisitic phenomena (e.g. Hurford et al. (1998)). One should resist the tendency to demonize generative grammar on the grounds that it seems to attribute discontinuity of capacity between humans and other animals. The emerging picture may be one in which human language acquisition has a strong evolutionary compotent with language specific developmental canalization that combines with more general aspects of cognition to generate language readiness (e.g. Batali (1994)). There is not enough evidence on ei-

ther side to conclusively say that human language acquisition capacity is primarily innate or primarily based on culture and general cognitive abilities. Language readiness of humans may also have some unexpected sources, combining the evolution of neurophysiology with other abilities, e.g. see the discussion of mirror neurons in monkey brain area F5 (which fire both when particular affordances are used in action by the animal or observed being used in actions of others) which is homologous to Broca's area in human for a proposed model of human language evolution (Rizzolatti and Arbib (1998); Arbib (to appear)).

The degree to which communicative systems are innate, subject to developmental variation and learning, and whether their learned aspects are mainly acquired via individual or social learning are often topics of heated debate. Of course, the degree to which and which aspects of such systems are innate will vary considerably from species to species.

5 Shared Meaning

Having rejected privileged agent-independent notions of semantics, meaning, signs, concepts and mappings, how is it possible to account for the fact that agents do in fact succeed in cooperation and communication? Does this not require us to resort to postulating external Platonist universals to which agents have at least limited access? No, it does not.

Similarities of experience between agents can account for the observed correspondences in the making of meaning. Agents sharing an environment, with similar sensory and actuator appartus, with similar bodies and needs will to lesser or greater degrees share modes of interaction with their world. Their Umwelts (worlds of experience around the agent) may correspond to lesser or greater degrees. The sharing of these features can be the substrate supporting similarity of sensory perceptions, similarity of actions, and needs (hence of what is useful for the agents). This can already account for innate similarities in the experience of meaning, and hence of the grounding for communication via similaries between the sender and receiver (Dautenhahn (1995); Nehaniv et al. (1999)). However, the sender and receiver of a signal may have radically different embodiments, such as echo-locating bats and their insect prey, dolphins and prey fish. In such cases, the signal may also result in transfer of information useful to one or both parties, but the meaning of the interaction is only shared in the sense that both parties take part in two different instances of semiosis in which there is overlap in the signal and possibly the signified vertices of the semiotic triangle.

In societies of interacting agents, there is an opportunity not only for the signs and signified to converge within distinct agents, but building on biological factors, there is also opportunity for further convergence by means of learning in the course of many interactions. This may result in a convergence of concepts, signifieds, and conventionalizations of signals into systems of signs. Moreover, the mappings, linking signs and signifieds, may also have the opportunity to converge. In this case, shared semiotic systems make communication more like – but still very distinct from the Platonistic idealization and simplification of denotational semantics with its "sign-meaning pairs". Beyond biologically innate or developmentally 'programmed' instances of such convergence, conventionalization of interaction via cultural transmission or social learning appears to be the only possible mechanism that can account for the emergence of such (shared) semiotic systems.

In interspecies interaction, parrots (Pepperberg (to appear)), chimpanzees, bonobos (Savage-Rumbaugh and Brakke (1996)) and bottlenosed dolphins (Herman and Austad (1996)) have all shown that they are capable of acquiring various components of human or human-constructed language-like communication systems, involving categories and reference, requests to satisfy intentions, and in the case of bonobos and dolphins, also the ability to understand, as evidenced by action in controlled experiments, syntactically complex imperatives, or again for dolphins, even notions of absence and abstract concepts such as simultaneity (tandem action) and imitation (Herman (to appear)). Social interaction (with humans) was a key feature in the animals' acquisition of these linguistic abilities.

M. Oliphant (Oliphant (to appear)) argues that as far as we know only humans have naturally occurring arbitrary symbolic reference. He shows that learning such arbitrary correspondences (between "meaning-symbol" pairs) is easily accomplished already using very simple artificial neural network models (e.g. using Hebbian learning), so computational capacity limitations of learning ability cannot be responsible for the observed apparent lack of learned arbitrary referential symbols in non-human animals. He speculates that this lack may be due to the difficulty in "observing meaning", i.e. other animals do not learn to communicate because of difficulty in "determining the meaning a signal is intended to convey." Meanwhile, humans use taxonomic categories, awareness of pragmatic context, reading the intent of the speaker, and human adults modify their utterances when speaking to younger children.

However, experiments with socially-mediated learning in (even differently embodied) robotic agents, show that at least acquisition corresponding labelling ("protowords") for similar external environments is possible via associative learning using temporal delays (Billard and Dautenhahn (1999)).

All of this suggests that shared meaning (corresponding processes of semiosis) requires shared experience in a social setting (or biologicial innate similarity). It is important in the social acquisition of sign systems that agents are allowed to attempt uses of communication to meet their own goals (e.g. intentions, homeostasis, transportation, feeding needs) rather than those of experimenter or other agents (Savage-Rumbaugh and Brakke (1996)). This is in accord with the notion that meaning depends on usefulness to the agents, and thus motivates the acquisition of the semiotic system, as when human children acquire human language.

6 Interaction Games

In this section, we will look inside communication and examine some of important features that are present in at least some forms of animal or human communication.

6.1 Language Games

Wittgenstein viewed natural language as comprised of myriad (and often very separate) language games in which language is employed in a particular contexts by participants in a particular manners. He constructed many examples of *language games* played according strict rules in his philosophical investigations (Wittgenstein (1968)) to gain insight into the nature of language and other topics. In each game participants (or, agents, if you like) use language to accomplish certain things in the world. Wittgenstein uses the word 'grammar' to describe the use of language or language components (whether natural, formal, or artificial) in carrying out particular tasks or activities. E.g. children singing 'Ring around the rosy, a pocket full of posies, Ashes, Ashes, we all fall down' when dancing in a circle holding hands; making a list of items to buy at a grocery store, and then checking them off the list as they are collecting into one's shopping basket; asking another person the time; yelling 'brick' or 'slab' at a construction site when asking another worker to bring the needed object. Many of Wittgenstein's examples include simple finite languages with strict rules of use, but the notion includes all ways in which natural language is employed.

Context is crucial in language games. When the rules of one game are applied in the context of another situation, interaction may fail, or we may produce in ourselves a sense of confusion or bewilderment. For example, the syntax of natural language allows us to say "Where is the book?", an ordinary question we might ask in trying to obtain an item. Since "the book" is a noun phrase, we might substitute another noun phrase such as "the universe" or "toothache" to create unusual questions, which seem meaningful since we can form them syntatically. Yet they are not part of our everyday life language games and so are not "grammatical" within these games. Similarly, since we can say "What happened before Thursday?", syntax allows us to say "What happened before time?". Much of philosophy begins with attempts to interpret such use of language outside the ordinary contexts of its uses in natural language games.

Real agents only play the language games that are useful for them. A statement like "This pen is blue" is never made about a pen that the speaker knows is red, unless there is a reason behind this. Examples games in which this could occur: the speaker wishes to deceive or manipulate others; the speaker is illustrating the possibility of counterfactuals (in doing philosophy - i.e. playing a philosophy language game).

6.2 Interaction Games

Generalizing Wittgenstein's notion of language games to non-linguistic realms, the author has described *interaction games* in which agents employ channels of sensing and actuation in some manner that is useful for them (Nehaniv (1999)). This is essentially the notion of language game, except that it has been minimally expanded so that it now easily applies to non-human animals and (robotic or software) agents. The notion of interaction games, includes animal communication and signalling (see below), and since the notion "useful" can be well-defined in terms of reproductive success or evolutionary terms, the identification and study of interaction games in the animal world provides part of the basis for a evolutionary continuity between humans and other animals. Such parsimony between explanations of human and animal features of interaction and communication is a theme of evolutionary psychology (Byrne and Whiten (1988)), cognitive ethology (Griffin, 1976, p. 102), or the study of animal minds (Griffin (1992); Jamieson and Bekoff (1996)).

6.3 Games Animals Play

Formalized signalling interactions are apparent in the natural behavior of many animals. In dogs a 'play bow' may precede what would otherwise appear to be aggressive or sexual behaviour (Bekoff (1977)). Marking a sequence by a preceding play bow tells the canid observer "what follows is play". Squids, cuttlefish and octopi employ elaborate signaling systems for attracting a mate, threatening rivals, hunting, confusing or frightening others and for camouflage. Chromatophores in the skin of many cephalopods allow them via fast neural control to alter their body

patterning, to signal to conspecifics or members of other species, even sending different signals to different observers viewing the animals from various perspectives (Moynihan (1985); Hanlon and Messenger (1996)). Squids can very quickly change from one display to another in a sequence. It is unclear whether and to what degree these changes are syntactically governed.

6.4 Comprehension / Production

Humans (and other animals or agents) may have differ capacity in comprehending as compared to producing communicative signals. Generally, but not always, ability to receive and interpret (parse or act on) communication is higher than the ability to produce the signals as evidenced in humans, apes, and dolphins (Herman and Austad (1996)).

6.5 Deixis

The indication of direction or directional reference to objects in language and interaction is called *deixis*. We see it in humans in deictic gaze (already present in prelinguistic infants) and also in words like "this" and "those".

Ants pheromones seem to have deictic qualities. And the use of honeybee dances to point in a sophisticated way that indicates both direction and distance is another. Despite what is sometimes asserted, the honeybees' dances do not refer only to sources of food, but may be employed also for other deictic functions such as the indication of desirable nesting sites (Griffin (1976)).

6.6 Reference, Categories, and Naming

Labelling particular objects, or categories of objects is a property of human language. More generally, not only objects, but attributes, actions, and relationships can be named with words. Categories can group together entities based on functional similarity, i.e. the fact that they require similar behavioural responses, or on syntactic similarly, i.e. a degree of interchangability between words of the same category in the structure of utterances (e.g. transitive verbs, animate singular nouns, etc.) How such categories might arise in humans and animals is unclear. But artificial neural network models in which the output is to behavioural selection rather than classification might lead to insight. Clustering into categories can thus arise via separability, or via association of objects with similar properties (i.e. similar to the agent perceiving them).

Reference for proper names (signals labelling unique items, places or individuals) is less of problem than is the origin of abstract nouns, classes, categories, verbals, attributes, and relation words.

6.7 Association vs. Predication

Hebbian learning and concept formation using artificial neural networks may be sufficient for describing the phenomenon of association, and even for some cases of action selection. Association is generally symmetric, but can be made asymmetric, e.g. through the use of temporal delay information. Predication is a particular type of asymmetric association, as in assertions that some entity has a property, and its weaker cousin modification, which is function of adjectives and adverbs, which are responsible for a kind of less marked predication in language. Grades of abstractness predication depend on the notion of category (e.g. entity with proper name or generic entity) and attributes (properties). There seems to be no evidence for natural occurring instances of predication in non-human animals. Why this is so remains to be explained. Predication may lie at the core of human syntax. Weaker versions of it seen in human language include topic comment constructions.

6.8 Discrimination Games

Pepperberg (to appear) presents evidence for prediction, attribute of properties to objects in African Grey parrots trained using a socially-based model rival technique. Apes can use attribute labels (e.g. Savage-Rumbaugh and Brakke (1996)) and bottlenosed dolphins demonstrate understanding of absence vs. presence of objects and distinguish possible vs. impossible requests in a syntactic command language used with them by human trainers (Herman and Austad (1996)). We can call games in which an agent must indicate or possibly even predicate that an entity has a property *discrimination games*. In many cases it is still unclear to what degree what is happening is like predication in human language.

This sort of interaction game is employed by Steels (1995) in experiments with software agents and robots. With possible referents given a priori in his model along with separation of sensory channels, individuals in the game attempt to refer to the same object in the environment. This goal of reference is built in, as is the notion of predication. Success in this game occurs if the predicate (given by the sender) uniquely determines the entity of which the predication is made to the recipient. Iterated playing of the the game leads to convergence of (proper) names labelling of entities, and of either spatial predicates that determine a third entity, or, alternatively, of predicates that constrain ranges of (sometimes several) feature values. Within each agent, phonetic symbols are associated to ranges of values in sensory channels. Communicative success

is the criterion each agent uses in deciding whether or not to revise their association of phonetic elements to labels for objects or for attributes. Although the models of (Steels (1995)) have built in capacities for reference and predication, the system does illustrate how conventions of labelling can arise in a population that has such capacities, even if the set of objects and attributes is open and changing.

Explaining how reference and predication could arise remains an open problem.

6.9 Following Games

In *following games* (employing learning by imitation), signals are employed to ensure the coordinated movement of teacher and student robots. Additionally, short binary string signals ('words') are emitted by the teacher as a function of its sensor values. By using an appropriate delay parameter (related to body length and speed of motion), the student comes associate the words with its own sensory experience in similar contexts. Thus the 'meaning' of the signals is acquired (Billard and Dautenhahn (1999)). Here the signals are from a small finite set, but the perceptions they are associated with need not be similar since the technique works even with agents having different body architectures.

7 Syntax

Syntax (rules of grammar) is often considered by linguists as being absolutely necessary for human like linguistic ability. Some precursors and features are the combination of symbols to yield new types of communicative acts not previously possible (Savage-Rumbaugh and Brakke (1996)), rule sets generating finite sets of possible signalling events, compositional or subcategorization structure, and recursion and combinatorial explosion in the number of possible communicative acts (see below).

7.1 Compositional Structure

The language used by Herman and Austad (1996) with dolphins had a strict word order in which target goals occur first, objects to be manipulated occur in second position, and actions occur last. While still finite (though extensible), this language has *compositional syntactic structure*: commands in the language take arguments whose role is determined by position. Allowing other marking (other than position) to indicate role would also yield compositional syntax.

Lexical items can take arguments (subcategorization), e.g. VP $\rightarrow$ V NP, a verb phrase may be constituted from a verb followed by a noun phrase as in $[_{VP}[$ eats $]_V[_{NP}$ the chocolate cake $]_{NP}]_{VP}$. Grammatically "the chocolate cake" is the direct object of the "eats". "Eats" has constituents or slots, including an object slot. The correspondence between the argument structure in syntax and semantics is also sometimes called 'compositionality' (e.g. Kirby (1999)), but this might more precisely be called homomorphic mapping or morphism or, more generally, a structure-preserving map (e.g. Goguen (1999)), i.e. the terms of logical form, syntatic representation, and phonetic form can be obtained via structure preserving correspondences. This is what Chomsky calls the 'projection principle' (Chomsky (1981),(Sells, 1985, p. 33)).

7.2 Recurison

When a lexical item subcategorizes for other items, it may be that by following a chain of such subcategorizations that it is possible to reach an another item of the original type. E.g. "I believe that you think ...", in such cases recursion is possible. Or in phrase structure rules

$$X \stackrel{*}{\rightarrow} \alpha X \beta,$$

where X is a non-terminal and α, β are some strings. More generally, the exponent growth in the number of generated strings can result when there are derivations of the form

$$X \stackrel{*}{\rightarrow} \alpha X \beta$$

with α and β non-empty or

$$X \stackrel{*}{\rightarrow} \alpha X \beta X \gamma.$$

Recursion and related exponential growth in generative capacity are extremely likely to arise in random sets of rules for context-free grammars.

Formal language theory, concerned with the description of sets of strings, provides convenient methods to describe such structure. Chomsky's *Syntactic Structures* show that (while English is not a context-free language) a context-free formal grammar can give an approximation of a fragment of English. The same holds for other human natural languages. The formalism works well for computer languages such as PASCAL, FORTRAN, C, etc., which are actually defined using such formalisms. Semantics of these languages is compositional in the sense that fixed meanings percolate up from leaf nodes in the parse tree of the language statement, and functions at intermediate nodes are applied to the node's constitutent argument list. (E.g., consider how an assigment statement like $X := C+5$ is parsed: the value of variable "C" and integer "5" are arguments to function "+", so that $C + 5$ comprises an expression evaluated by applying addition to these arguments; while the assignment operator ":=" takes a variable and expression as its arguments, evaluates the expression and assigns the result to the variable X.).

First-order and higher-order logic formulae are similarly constructed using context-free grammars. Truth values of formulae in the languages determined by these grammars are similarly determined (with respect to a particular structure or "world of discourse" over which the interpretation is made) by recursion application of rules which finally reduce to the assignment of truth values to the equality of terms and the truth values of predicates. Rather than inducing well-defined operations in a computer, the interpretation of a logical formula over a structure returns either "true" or "false". Once the structure and rules of interpretation have been thus specified, all observers will assign the same truth value to each formula.

Predication is built into the ediface of formal logic. Constituent argument structure ("compositionality") is built into the formalism of first-order logic and into the structure of programming languages, and other formalisms. These properties were abstracted from natural language by logicians and mathematicians. They have been codified and standardized in such a way that someone using them is able to 'escape from context', i.e. knowledge expressed in such formulae is an example of what Bruno Latour (Latour (1987) has called an 'immutable mobile', knowledge that can be reused in other contexts when applied according to certain general procedures or rules. Joesph Goguen has called this 'dry' information, as opposed to 'wet' information which cannot be interpreted outside its particular original grounded, embedded, situated context. Note that there are degrees of dryness and wetness, or in Latour's terms, degrees of mobility. For example, a cake recipe, is a partly formal but reusable piece of information somewhere in the middle of the wet-dry continuum.

These formal properites compositionality (argument structure, subcategorization) and semantics of predication are thus very well-supported by the tools of computer science and formal grammars. It is very easy to describe compositional formal language systems and associated semantics using these tools. That is exactly what the tools were developed for. Tools such as context-free grammar (Backus-Naur form), phrase structure grammars, denotational semantics, programming languages, etc., abstract from structure of natural human language and also 'clean-up' the embeddedness ('wetness') increasing the mobility of knowledge (well-definedness of truth values of formulae when interpreted over structures, portability of software, etc.)

It should therefore come as no surprise if we observe the "emergence" of predication or compositionality or of recursion in models of the evolution of communication and evolution of language which formulate their grammars using tools of context-free grammar or subcategorization in argument structure: That latter were constructed to facilitate the former.

8 Random drift: "Diversity" and "Convergence"

In repeated stochastic sampling of a population, the distribution in the sample is unlikely to exactly match the distribution of characters in the population. This phenomenon is well-known in statistics, where large sample sizes and confidence intervals are used to limit and quantify the likely effects of sampling error (Freedman et al. (1997)). In evolutionary genetics (Maynard Smith (1989); Sigmund (1993); Roughgarden (1996); Schmitt and Nehaniv (1999)) repeated sampling of a finite population (an all biological populations are finite) results in genetic drift of the inherited traits (independent of natural selection and variation due to mutation) towards random but uniform values. Explicit bounds on the rate of convergence due to genetic drift in iterated random sampling with or without the action of selective pressure have been calculated (see the above references). It is a mathematical theorem, that under very general circumstances, e.g. in the absence of mutation, a fixed-size finite population subject to any operators of fitness selection and with or without sexual recombination will converge (with probability 1) to a population of individuals all having the same genotype. Moreover, this is even true, if for instance, what is transmitted is not called 'genotype' but is e.g. a 'meaning-symbol' map acquired from observation of other agents' use of 'language'. This is all that is behind the so-called 'emergence of a common language' in some computational models. Sometimes such random drift convergence has be given the name 'self-organization'.

Convergence can be prevented by the introduction of random variation in the course of reproduction (e.g. the random resetting of bits in a genetic algorithm). These mechanisms by themselves explain much of what is seen e.g. in the results of Arita and Koyama (1998) on so-called "linguistic diversity".

Cases of random drift and drift combined with selection and variation are seen, for example, in the studies of Arita and Koyama (1998) at a genetic level for individuals defined by meaning-symbol pairs, of Steels (1995) in which entities consist of sets of meaning-symbol pairs but modify themselves (selection and variation) based on communicative success, and of (Hashimoto and Ikegami (1995); Steels (1998); Kirby (1999)) in which individuals can at least roughly be viewed as grammars, i.e. populations of sets of rules.

'Emergence' and 'self-organization' are terms by experimenters to describe phenomena which surprise them and for which they can offer no detailed explanation. Minsky has argued that use of the word 'emergence' should make one suspicious that not enough effort has been made in finding explanatory mechanisms (Minsky (1996)). If the criterion for emergence is one of surprising the investigators, then the notion

is clearly very much observer-dependent, in such a formulation of little value to science. However, emergence can be defined in a more formal way in terms of a rigorous mathematical definition of complexity as complexity increase in the extreme upper range of certain bounds on complexity increase (for one-way interactions) or greater increase (for interaction with feedback), see (Nehaniv and Rhodes (2000)).

9 Building the Solutions In

We have seen some evidence that simulation models without evolution of innate language ability can be put forward for possible explanatory mechanism of aspects of language or communication evolution. Steels' discrimination games (Steels (1995)) have also been extended to games in which not only phonemic labels, but constraints on ordering or introduced to model evolution of syntax (Steels (1998)). In the former predication and reference were built in to the agents, in the latter subcategorization frames are built in, i.e. compositionality is assumed, although not its particular realization under a mapping to 'surface structure'. Kirby (1999) starts with a space of privileged meanings that are compositional and recursive, and using context-free formalisms to acquire grammars which define structure-preserving maps from 'meanings' to 'utterances'; in this setting he shows that the bottleneck of learning (and certain generalizing variation operations) leads over time to increasingly generic context-free grammars that preserve structure of the external 'meaning' space. Hashimoto and Ikegami (1995) show that social factors can determine the communicative success of grammar using agents that play a game of generating and parsing abstract utterances. Subjacency, a structural constraint on argument chains in determining reference in universal grammar (e.g. (Sells, 1985, p. 48)) can probably be shown to arise once context-free like rules are employed in compositional syntax. The origin and maintenance of syntactic phenomena such as deixis, predication, compositionality, and grammars can still be considered wide open problems.

Innate language acquisition devices and language readiness (either neurophysiological, cognitive, or cultural) have been proposed but yet not demonstrated as sufficient to account for human linguistic capacities (Chomsky (1968); Pinker and Bloom (1990); Arbib (to appear); Hurford et al. (1998)). We expect a crucial role for social factors and interaction, at the level of individual development and in evolving populations or societies of agents.

References

Michael A. Arbib. The mirror system, imitation, and the evolution of language. In K. Dautenhahn and C. L. Nehaniv, editors, *Imitation in Animals and Artifacts*. MIT Press, to appear.

Takaya Arita and Yuhiji Koyama. Evolution of linguisitic diversity in a simple communication system. *Artificial Life*, 4(1):109–124, 1998.

J. Batali. Innate biases and critical periods. In Rodney Brooks and Patti Maes, editors, *Artificial Life IV*, pages 160–171. MIT Press, 1994.

Marc Bekoff. Social communication in canids: Evidence for the evolution of stereotyped mammalian display. *Science*, 197:1097–1099, 1977.

D. Bickerton. *Language and Species*. Chicago, 1990.

A. Billard and K. Dautenhahn. Experiments in learning by imitation - grounding and use of communication in robotic agents. *Adaptive Behavior*, 7 (3/4), 1999.

Jack W. Bradbury and Sandra L. Vehrencamp. *Principles of Animal Communication*. Sinauer, 1998.

J. Bruner. The narrative construction of reality. *Critical Inquiry*, 18(1):1–21, 1991.

Gordon M. Burghardt. Ontogeny of commuication. In Thomas A. Sebeok, editor, *How Animals Communicate*. Indiana University Press, 1977.

Richard Byrne and Andrew Whiten, editors. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford, 1988.

Noam Chomsky. *Language and Mind*. Harcourt, Brace and World, 1968.

Noam Chomsky. *Reflections on Language*. Pantheon Books, 1975.

Noam Chomsky. *Lectures on Government and Binding*. Foris Publications, 1981.

Charles Darwin. *The Expression of the Emotions in Man and Animals*. 1872.

Kerstin Dautenhahn. Getting to know each other – artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, 16:333–356, 1995.

David Freedman, Robert Pisani, and Roger Purves. *Statistics*. W.W. Norton & Co., 3rd edition, 1997.

Joseph A. Goguen. An introduction to algebraic semiotics, with application to user interface design. In *Computation for Metaphors, Analogy, and Agents*, volume 1562, Lecture Notes in Artificial Intelligence, pages 242–291, 1999.

Donald R. Griffin. *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience.* The Rockefeller University Press, 1976.

Donald R. Griffin. *Animal Minds.* Chicago, 1992.

Roger T. Hanlon and John B. Messenger. *Cephalopod Behaviour.* Cambridge University Press, 1996.

Takashi Hashimoto and Takashi Ikegami. Evolution of symbolic grammar system. In F. Morán, A. Moreno, J. J. Merolo, and P. Chacón, editors, *Advances in Artificial Life (3rd European Conference on Artificial Life, Granada, Spain, June 1995)*, volume Lecture Notes in Artificial Intelligence, 929, 1995.

Louis M. Herman. Vocal, social, and self imitation by bottlenosed dolphins. In K. Dautenhahn and C. L. Nehaniv, editors, *Imitation in Animals and Artifacts.* MIT Press, to appear.

Louis M. Herman and Steven N. Austad. Knowledge acquisition and asymmetry between language comprehension and production: Dolphins and apes as general models for animals. In Marc Bekoff and Dale Jamieson, editors, *Readings in Animal Cognition*, pages 289–306. MIT Press, 1996.

James Hurford, Chris Knight, and Miceal studdert Kennedy, editors. *Approaches to the Evolution of Language: Social and Cognitive Bases.* Cambridge University Press, 1998.

Dale Jamieson and Marc Bekoff. On the aims and methods of cognitive ethology. In Marc Bekoff and Dale Jamieson, editors, *Readings in Animal Cognition*, pages 65–77. MIT Press, 1996.

Simon Kirby. Learning, bottlenecks, and infinity: a working model of the evolution of syntactic communication. In K. Dautenhahn and C. L. Nehaniv, editors, *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*, pages 55–63. Society of the Study of Artificial Intelligence and the Simulation of Behaviour, 1999.

Bruno Latour. *Science in Action.* Harvard, 1987.

Bruce MacLennan. Synthetic ethology: An approach to the study of communication. In *Artificial Life II.* Addison Wesley, 1992.

John Maynard Smith. *Evolutionary Genetics.* Oxford University Press, 1989.

John Maynard Smith and Eörs Szathmáry. *The Major Transitions in Evolution.* W.H. Freeman, 1995.

Marvin Minsky. Plenary address at Artificial Life V, the 5th international workshop on the synthesis and simulation of living systems. 1996.

Martin Moynihan. *Communication and Non-Communication by Cephalopods.* Indiana University Press, 1985.

Chrystopher L. Nehaniv. Meaning for observers and agents. In *IEEE International Symposium on Intelligent Control/Intelligent Systems and Semiotics, ISIC/ISAS'99 (September 15-17, 1999 Cambridge, Massachusetts)*, pages 435–440, 1999.

C. L. Nehaniv, K. Dautenhahn, and M. J. Loomes. Constructive biology and approaches to temporal grounding in post-reactive robotics. In *Sensor Fusion and Decentralized Control in Robotics Systems II (September 19-20, 1999, Boston, Massachusetts), Proceedings of SPIE Vol. 3839*, pages 156–167, 1999.

Chrystopher L. Nehaniv and John L. Rhodes. The understanding and evolution of biological complexity from an algebraic perspective. *Artificial Life*, 6 (1), 2000.

Michael Oliphant. Rethinking the language bottleneck: Why don't animals learn to communicate? In K. Dautenhahn and C. L. Nehaniv, editors, *Imitation in Animals and Artifacts.* MIT Press, to appear.

Charles S. Peirce. *Collected Papers, Volume 2: Elements of Logic.* Harvard, 1995.

Irene M. Pepperberg. Allospecific referential speech acquisition in grey parrots (*Psittacus erithacus*): Evidence for multiple levels of avian vocal imitation. In K. Dautenhahn and C. L. Nehaniv, editors, *Imitation in Animals and Artifacts.* MIT Press, to appear.

S. Pinker and P. Bloom. Natural language and natural selection. *Brain & Behavioral Sciences*, 13: 707–784, 1990.

G. Rizzolatti and M. A. Arbib. Language within our grasp. *Trends in Neurosciences*, 21(5):188–194, 1998.

Jonathan Roughgarden. *Theory of Popultion Genetics and Evolutionary Ecology: An Introduction.* Prentice Hall, 1996.

Sue Savage-Rumbaugh and Karen E. Brakke. Animal language: Methodological and interpretative issues. In Marc Bekoff and Dale Jamieson, editors, *Readings in Animal Cognition*, pages 269–288. MIT Press, 1996.

L. M. Schmitt and C. L. Nehaniv. The linear geometry of genetic operators, with applications to the analysis of genetic drift and genetic algorithms using tournament selection. In *Mathematical and Computational Biology*. American Mathematical Society, 1999.

Peter Sells. *Lectures on Contemporary Syntactic Theories*. Center for the Study of Language and Information, Stanford, 1985.

R.M. Seyferth and D . L. Cheney. Vocal development in vervet monkeys. *Animal Beahviour*, 34:1640–1658, 1986.

C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1963.

Karl Sigmund. *Games of Life: Explorations in Ecology, Evolution and Behaviour*. Penguin, 1993.

W. John Smith. *The Behavior of Communicating: An Ethological Approach*. Harvard, 1977.

W. John Smith. Communication and expectations: A social process and the operations it depends upon and influences. In Marc Bekoff and Dale Jamieson, editors, *Readings in Animal Cognition*, pages 243–255. MIT Press, 1996.

Luc Steels. A self-organizing spatial vocabulary. *Artificial Life*, 2:315–332, 1995.

Luc Steels. The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103:1–24, 1998.

Ludwig Wittgenstein. *The Blue and Brown Books*. Harper & Brothers, 1958.

Ludwig Wittgenstein. *Philosophical Investigations (Philosophische Untersuchungen (German with English Translation)*. Basil Blackwell, 3rd edition, 1968.

Imitation and Reinforcement Learning in Agents with Heterogeneous Actions

Bob Price*;Craig Boutilier†
*Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada V6T 1Z4
†Department of Computer Science, University of Toronto, Toronto, ON, Canada M5S 3H5
price@cs.ubc.ca;cebly@cs.toronto.edu

Abstract

We study the problem of accelerating reinforcement learning through the observation and *implicit imitation* of expert agents (mentors) acting in the same domain. In this paper, we consider problems that arise when the learner and mentor have heterogeneous actions. We extend an earlier implicit imitation model to allow for feasibility testing (determining whether a specific mentor action can be duplicated) and repair (discovering a "plan" that simulates a mentor's trajectory) and demonstrate empirically that both of these components allow learning agents to learn much more readily than standard RL agents and implicit imitation agents without these extended capabilities.

1 Introduction

Cooperative multiagent systems rely on shared models and communication to coordinate their actions in a common environment. While many researchers have examined explicit communication systems, we have argued (as have others) that implicit communication techniques such as imitation increase the range of applications for multi-agent systems and pose interesting cognitive models of interaction in agent societies (Dautenhahn, 1995; Price & Boutilier, 1999). In an imitation model with implicit communication, agents can learn from others without communicating an explicit context for the applicability of a behaviour (Bakker & Kuniyoshi, 1996); without the need for a pre-existing communication protocol; in competitive situations where agents are unwilling to share information; and even when other agents are unwilling to fulfill a teacher role. The ability of imitation to effect skill transfer between agents has been demonstrated in a range of domains (Atkeson & Schaal, 1997; Billard & Hayes, 1997; Hayes, 1994; Kuniyoshi et al., 1994; Mataric, 1998; Mitchell et al., 1985; Utgoff & Clouse, 1991). These domains, however, have primarily dealt with agents imitating other agents with essentially the same action set as themselves. Our goal is to extend the benefits of imitation to situations in which the action capabilities of agents in the environment differ from one another.

In previous work we showed that *implicit imitation* can improve a reinforcement learner's effectiveness by allowing it to take advantage of the knowledge implicit in observations of more knowledgeable agents (Price & Boutilier, 1999). Though we made no assumption that the learner shared the same objectives as the mentors, we did rely crucially on the fact that actions were *homogeneous*: every action taken by a mentor corresponded to some action available to the learner. In this paper, we relax this assumption and introduce several mechanisms that allow acceleration of RL in presence of *heterogeneous actions*. Specifically, we introduce two notions: *action feasibility testing*, which allows the learner to determine whether a specific mentor action can be duplicated; and *k-step repair*, in which a learner attempts to determine whether it can approximate the mentor's trajectory. Both of these concepts are used to modify the influence that mentor observations have on the learner's estimate of its own value function.

Our work can be viewed (loosely) as falling within the formal imitation framework proposed by Nehaniv and Dautenhahn (1998), who propose viewing imitation as the model-based process of constructing mappings between states, actions, and goals of different agents (see also the abstraction model of Kuniyoshi at al. (1994)). However, key differences include the fact that we assume that state-space mappings are given, that the mentor's actions are not directly observable, that the objectives (goals) of the mentor and learner may be different, and that our environments are stochastic. Furthermore, we do not require that the learner explicitly try to duplicate the behavior of the mentor. In this way, our model, like (Atkeson & Schaal, 1997) differs from "following" and "demonstration" models often used in robotics (Hayes, 1994; Mataric et al., 1998). However, the repair strategies we invoke do bear some relation to "following" models.

2 Homogeneous Actions

In *implicit imitation* (Price & Boutilier, 1999), we assume two agents, a *mentor* m and an *observer* o, acting in a fixed environment.[1] We assume the observer (or learner), o, is learning to control an MDP with states S, actions A_o and reward function R_o. We use $\Pr_o(t|s,a)$ to denote the probability of transition from state s to t when action a is taken.

[1]The extension to multiple mentors is straightforward (Price & Boutilier, 1999).

The mentor too is controlling an MDP with the same underlying state space (we use A_m, R_m and $\Pr_m$ to denote this MDP).

We make two assumptions: the mentor implements a deterministic[2] stationary policy π_m, which induces a Markov chain $\Pr_m(t|s) = \Pr_m(t|s, \pi_m(s))$ over S; and for each action $\pi_m(s)$ taken by the mentor, there exists an action $a \in A_o$ such that the distributions $\Pr_o(\cdot|s, a)$ and $\Pr_m(\cdot|s)$ are the same. This latter assumption is the *homogeneous action assumption* and implies that the learner can duplicate the mentor's policy. We do not assume that the learner knows *a priori* the identity of this action a (for any given state s), nor that the learner *wants* to duplicate this policy (the agents may have different reward functions). Since the learner can observe the mentor's transitions (though not its actions directly), it can form estimates of the mentor's Markov chain, along with estimates of its own MDP (transition probabilities and reward function).

We define the *augmented Bellman equation* as follows:

$$V(s) = R_o(s) + \gamma \max \left\{ \max_{a \in A_o} \left\{ \sum_{t \in S} \Pr_o(t|s,a) V(t) \right\}, \sum_{t \in S} \Pr_m(t|s) V(t) \right\} \quad (1)$$

This is the usual Bellman equation with an extra term added, namely, the second summation, denoting the expected value of duplicating the mentor's action $\pi_m(s)$. Since this (unknown) action is identical to one of the observer's actions, the term is redundant and the augmented value equation is valid. Furthermore, under certain (standard) assumptions, we can show that the estimates of the model quantities will converge to their true values; and an *implicit imitation learner* acting in accordance with these value estimates will converge optimally under standard RL assumptions.[3] More interesting is the fact that by acting in accordance with value estimates produced by augmented Bellman backups, an observer generally converges much more quickly than a learner not using the guidance of a mentor. As demonstrated in (Price & Boutilier, 1999), implicit imitators typically accumulate reward at a higher rate earlier than standard (model-based) RL-agents, even when the mentor's reward function is not identical to the observer's.

At states the mentor visits infrequently (because they are not traversed in the optimal policy), the learner's estimates of the mentor's Markov chain may be poor compared to the learner's own estimated action models. In such cases, we would like to suppress the mentor's influence. We do this by using model confidence in augmented backups. For the mentor's Markov chain and the observer's action transitions, we assume a Dirichlet prior over the parameters of each of these multinomial distributions. From sample counts of mentor and observer transitions, the learner updates these distributions. Using a technique inspired by Kaelbling's (1993) interval estimation method, we use the variance in our estimated (Dirichlet) distributions for the model parameters to construct lower bounds on both the augmented value function incorporating the mentor model and an unaugmented value function based strictly on the observer's own experience. If the lower bound on the augmented value function is less than the lower bound on the unaugmented value function, we suppress the influence of the mentor and use an unaugmented Bellman backup.

[2] We could generalize the algorithm to stochastic policies

[3] We assume that an appropriate exploration strategy is being used and that it is influenced by estimated value; i.e., the learner is more likely to choose actions with higher estimated value

3 Implicit Imitation with Heterogeneous Actions

When the homogeneity assumption is violated, the implicit imitation framework described above can cause the learner to perform very poorly. In particular, if the learner is unable to make the same state transition (or a transition with the same probability) as the mentor at a given state, it may drastically overestimate the value of that state. Furthermore, there is no mechanism for removing the influence of the mentor's Markov chain on value estimates—the observer can be extremely (and correctly) confident in the mentor's model. The problem lies in the fact that the augmented Bellman backup is justified by the assumption that the observer can duplicate *every* mentor action.

To overcome this difficulty, we propose two techniques that allow observers to retain the guidance of mentors, but suppress the guidance when it is apparent that it is misleading. The more fundamental of these, but in some sense the more straightforward, is action feasibility testing: intuitively, when the learner is sure that it cannot duplicate the mentor's action at a given state, it suppresses the effect of augmented backups at that state (reverting to standard Bellman backups).[4] The technique is simple and eliminates the "lockup" effect sometimes observed in the basic implicit imitation framework when agents have differing capabilities. Unfortunately, this can sometimes cause useful guidance (in the form of higher value estimates) to be "cut off" in certain cases where that guidance would be useful. Specifically, when the learner can "repair" the mentor's trajectory by finding a (short) sequence of its own actions that leads to the same state as the infeasible action, the value guidance is likely appropriate. For this reason, we introduce the notion of *k-step repair* and a method for deciding when to allow mentor guidance to persist at a state despite the infeasibility of the mentor's action for the observer.

3.1 Action Feasibility Testing

The Dirichlet distributions used by our model-based RL-agent can be used to find the variance associated with a transition probability estimate. This variance can be used to test the feasibility of a mentor's action. To examine a simple case, suppose that there are only two successor states, t and u, for a specific action a_o taken at s (thus we estimate only one probability $\Pr_o(t|s, a_o)$). Further suppose that the mentor's action is similarly restricted and the mentor's Markov chain at that state is modeled by $\Pr_m(t|s)$. We

[4] The decision is binary; but we could envision a smoother decision criterion that measures the extent to which the mentor's action can be duplicated. We do not pursue this generalization here

could test statistically whether the two actions, a_o and the mentor's action, are the same by performing a difference of means test using the hypothesis that the mean probability of getting to state t is the same for both actions. Under this hypothesis we use the pooled variance of the two statistics which is computed by weighting the variances according to the number of samples used for each statistic.

$$\frac{Pr(t|a_1) - Pr(t|a_2)}{\sqrt{\frac{n_1(t|a_1)Var(t|a_1)+n_2(t|a_2)Var(t|a_2)}{n_1(t|a_1)+n_2(t|a_2)}}} > Z_{\alpha/2} \quad (2)$$

The Dirichlet distribution is highly non-normal for small sample size, so we construct our test criterion $Z_{\alpha/2}$ using the Tchebychev inequality, which is valid for any distribution. When the value of the left side of Equation 2 is greater than the right, we conclude that the actions are different and that there is no point in having the observer attempt to duplicate the mentor.

Generally, however, we will have a number of possible outcomes for an action (not just two) so we must perform a multivariate difference of means test. For well-behaved distributions (e.g., normal) there exist multi-variate difference of means tests (Scheffe, 1959). The work specific to multivariate testing of Dirichlet or generalized beta distributions assumes a sufficient number of samples to make the bounds computed reasonably tight (Goodman, 1965). A second method applicable to multivariate Dirichlet distributions is the Bonferroni Test (Seber, 1984) which allows one to construct a multivariate test from univariate components. It makes no assumptions about normality or independence and in comparison with techniques like (Goodman, 1965), it has been shown to give good results in practice (Mi & Sampson, 1993). Since it is also easy to implement and fast to compute, we employed the Bonferroni method in our implementation.

The idea behind the Bonferroni test is to perform a multivariate hypothesis test by conjoining several single variable tests. More generally, we might have a set of r specific hypotheses $E_1, E_2, \ldots, E_r$ that we wish to test simultaneously. Let $\bar{E}_i$ be the complementary hypothesis of E_i. The Bonferroni inequality tells us:

$$Pr\left[\bigcap_{i=1}^{r} E_i\right] \geq 1 - \sum_{i=1}^{r} Pr\left[\bar{E}_i\right]$$

Thus we can obtain a probability of α for the joint hypothesis $\bigcap_{i=1}^{r} E_i$ by testing each of the r complementary hypotheses $\bar{E}_i$ at α/r. The individual hypotheses E_i do not have to be independent. In testing for action equivalence, our individual hypotheses correspond to tests to see if the transition probability to a particular successor state is the same for both actions and the joint hypothesis is that all successor state transition probabilities are the same for both actions. We therefore set r to be the number of successor states.

To summarize, we test the distribution of successor states for the mentor's unknown action against the distribution of successor states for *each* of the observer's actions using a Bonferroni test. If all of the observer's experience-based action models are rejected, then it concludes that the mentor's action is infeasible and the influence of the model derived from mentor observations is suppressed. The algorithm is summarized in Figure 1.

```
FUNCTION feasible(m,s) : Boolean
    FOR each a_i in A_o do
        allSuccessorProbsSimilar = true
        FOR each t in successors(s) do
            μ_Δ = Pr_o(t|s,a) − Pr_m(t|s)
            z_Δ = μ_Δ / sqrt(var_o(t|s,a) + var_m(t|s))
            IF z_Δ > z_{α/r}
                allSuccessorProbsSimilar = false
        IF allSuccessorProbsSimilar
            return true
    return false
```

Figure 1: Action Feasibility Testing

3.2 k-Step Repair

Even if an observer cannot duplicate a mentor's primitive action at a particular state, guidance from the mentor may still be useful if the trajectory of the mentor through the state space is broadly "similar" to a feasible trajectory for the observer. We can capture this notion of "similarity" by augmenting feasibility testing with a device that encourages the learner to find these "similar" trajectories.

For example, suppose the observer is at state s and the mentor has been observed to make the transition from state s to state t, to state u enough times that the observer's estimates of $\Pr_m(t|s)$ and $\Pr_m(u|t)$ are very confident (see Figure 2). Suppose also, that state u is a highly rewarding state for both the mentor and observer. On the basis of these observations, the observer assigns a high value to $V(t)$ and $V(s)$ and is thereby encouraged to move toward these states during exploration. But suppose that after some time the mentor's action at state t is judged to be infeasible (e.g., there is an obstacle navigable by the mentor but not the learner). Unless the observer has embarked on sufficient exploration in the area to discover an alternate path from s to u (e.g., through t') before the judgment, the value of state s will plunge immediately. This in turn eliminates the observer's future motivation to move towards state s and explore local alternatives from that point. If, however, the observer assumes by default that it has a roughly similar trajectory to that of the mentor, it may persist in backing up value from t to s in the belief that it will be able to discover a "local" path or *bridge* from s to u.

Intuitively, a bridge is a "short" feasible path which bridges the gap in the value function due to an infeasible action. It starts on the mentor's trajectory in the state where the observer cannot duplicate the mentor's action and then navigates around the infeasible transition before ending on a state also on the mentor's trajectory but downstream of the infeasible transition. Such bridges can provide important guidance in cases where the value at a state (as defined by the augmented Bellman backup) is determined by the mentor's action rather than the learner's own actions. At such states, value estimates drop drastically as soon as the mentor's action is discovered to be infeasible unless a bridge has been discovered.

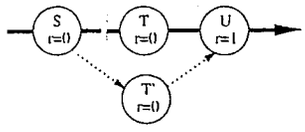

Figure 2: Prior Guidance

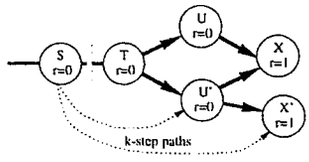

Figure 3: Reachability

We note that bridges are often formed naturally in the imitation model as formulated thus far. Frequently, the random exploration of the agent in its attempts to duplicate the mentor's path will cause it to sample states and actions in the general vicinity of the mentor's path. When an action on the mentor's path is judged infeasible, the alternative paths through the partially-explored—but until now unattractive—vicinity of the mentor become worth checking and thereby form bridges. In more difficult problems where there is little exploration significantly off of the mentor's trajectory, the background exploration of the agent will be insufficient to provide bridges.

A second source of bridges comes from the grid-world domain. Given a uniform prior over possible action effects,[5] each state is judged initially to be "reachable" with nonnegligible probability from states in its neighborhood. When a situation occurs (as described above) in which the mentor's action at s is deemed infeasible, the learner's value estimate $V(s)$ drops. However, this drop is often mitigated by the "flow" of value around the obstacle through neighboring states (e.g., t'). The use of uniform priors often seems to help this process along. This will encourage the learner to persist in exploring this neighborhood—thus, if a feasible bridge exists, it is likely to be found fairly early.

Prior guidance is not a reliable means of discovering bridges however. The combined effect of discounting and the small prior probability of state transitions cause the magnitude of value to decrease very rapidly with the length of the trajectory along which it is being backed up. Any negative rewards present can easily drown out small values. Thus at states significantly distant from s, the value gradient is unlikely to point toward s in a significant way. We therefore consider a more explicit means of encouraging exploration in the area. Our *k-step repair strategy* initiates explicit searches for bridges, specifies criteria for detecting their formation and caches the existence of a bridge in order to eliminate the need to check for it in the future.

k-step repair uses reachability analysis (based on the learner's current domain model) to test for the existence of a bridge. Consider the situation in Figure 3. When the learner first discovers that that the mentor's action at state s is infeasible, it undertakes a search for an existing bridge.

[5]We exploit local topology in our grid world experiments, so that a state is connected by any action *a priori* to its eight neighbours and to itself.

Let a *bridge termination state* be any state on the mentor's trajectory within the k steps following state with no feasible mentor action, s. The algorithm only steps along mentor transitions with greater than prior probability. The observer now searches for a bridge, also k steps long which starts at state s, follows only observed, feasible transitions and ends in a bridge termination state. Because only feasible transitions are considered, misleading priors do not have undue influence. If a bridge is found, the mentor's influence is ignored at state s as value should already be "flowing" back through the existing bridge. We flag the state as bridged so that we will not have to perform the bridge test again.

If a bridge is not found, however, we do not immediately suppress the mentor's influence at this state. Intuitively, we keep value flowing back to encourage the observer to come to the state with an infeasible action and explore the local neighbourhood before discounting the mentor's influence. If imitation is sensible in a given domain, we expect that it will be reasonable to assume that the path can be repaired by a short search of k-steps. The search is performed by a k^2-step random walk (in our 2-D grid worlds), which on average explores locations out to k-steps from the starting point (but not all locations up to k-steps away from s (Weisstein, 1996)). If during this walk the observer encounters a bridge termination state, we set the bridge-found flag for the originating state and suppress the value backup over the infeasible transition.[6] Attempts to discover bridges (as long as a bridge remains undiscovered) are performed n times (i.e., n visits to state s). During this time, suppression of the mentor's influence is itself suppressed. After n random walks, no more attempts are made, and the mentor's influence at state s is suppressed once and for all.[7] We note that k-step repairability could be developed into a measure of similarity between agents. The measure could be used to decide when it is worthwhile to attempt the repairs that would be required to imitate a given mentor.

Feasibility and k-step repair can be easily integrated into the existing imitation framework. The complete decision procedure appears in Figure 4. As in the original model, we first check to see if the observer's experience-based calculation for the value of the state has a better lower bound than the mentor-based calculation; if so, then the observer uses its own experience-based calculation. Otherwise, we check to see if the observer has a sufficient number of samples of its own behaviour to perform an action feasibility test. If not, we assume by default that the action taken by the mentor is feasible for the observer. This assumption will cause no permanent harm, as an error can only increase the value of the state which will in turn cause the observer to explore the state and increase the number of experience-based samples it has for this state. We cur-

[6]There is no guarantee that executing the *stochastic* action required to form a bridge will actually form the bridge on a given trial. Even if a bridge is discovered, there is no guarantee that it is optimal, but in our problems any bridge will increase the attractiveness of state s.

[7]One can determine "suitable" values for n using assumptions about the state space structure and noise level of actions. E.g., $n > 8k - 4$ seems suitable in an 8-connected grid world with low noise. We note that indiscriminantly large values of n can reduce performance below that of non-imitating agents.

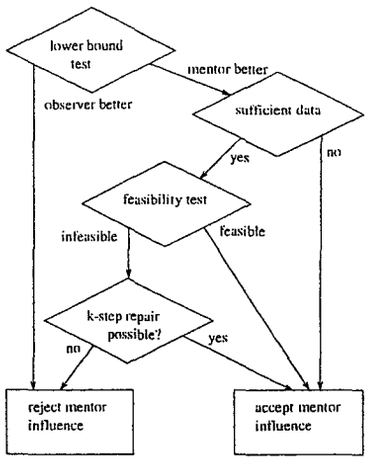

Figure 4: Implicit Imitation with Feasibility Tests

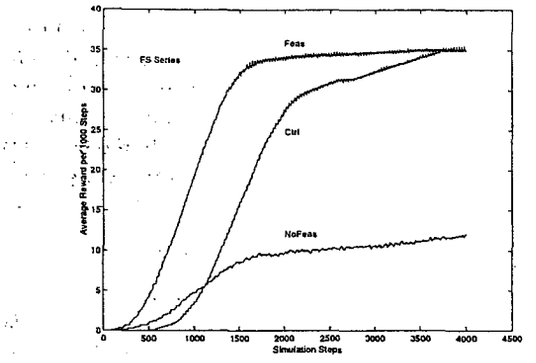

Figure 5: Utility of Feasibility Testing

rently use a threshold of 5 samples.

If there is sufficient data, we perform the action feasibility test. If the mentor's action is feasible, then we accept the value calculated using the mentor-observations based value function. If the action is infeasible we check to see if it is possible to do more bridging. The test checks two qualities of the state: If a bridge is already built then bridging is unnecessary. If we have exhausted our threshold for bridging attempts we say that it is impossible. In either case, no bridging actions are necessary so we can dispense with mentor guidance and use the observer's own experience-based calculations. If bridging is still possible then we delay suppression of mentor influence so that the augmented value function will guide the agent to the bridge building states and a repair can potentially be made.

4 Empirical Demonstrations

In this section, we empirically demonstrate the utility of feasibility testing and k-step repair and show how the techniques can be used to surmount both differences in actions between agents and small local differences in state-space topology. The problems here have been chosen specifically to demonstrate the necessity and utility of both feasibility testing and k-step repair. As space is limited here, we will refer to (Price & Boutilier, 1999) for a discussion on how the gains due to imitation increase with problem size and qualitative difficulty.

Our first experiment shows the necessity of feasibility testing in implicit imitation when agents have heterogeneous actions. In this scenario, all agents must navigate across an obstacle-free, 10-by-10 gridworld from the upper-left corner to a goal location in the lower-right. The agent is then reset to the upper-left corner. The first agent is a mentor with the "NEWS" action set (North, South, East and West movement actions). The mentor is given an optimal stationary policy for this problem. We study the performance of three learners, each with the "Skew" action set (N, S, NE, SW) and unable to duplicate the mentor exactly (e.g., duplicating a mentor's E-move requires the learner to move NE followed by S). The first learner employs implicit imitation *with* feasibility testing, the second uses imitation *without* feasibility testing, and the third control agent uses no imitation (i.e., is a standard RL-agent). All agents experience limited stochasticity in the form of a 5% chance that their action will be randomly perturbed. As in (Price & Boutilier, 1999) the agents use model-based reinforcement learning with prioritized sweeping (Moore & Atkeson, 1993).

The effectiveness of feasibility testing in implicit imitation can be seen in Figure 5. The horizontal axis represents time in simulation steps and the vertical axis represents the average number of goals achieved per 1000 time steps (averaged over 10 runs). We see that the imitation agent with feasibility testing converges much more quickly to the optimal goal-attainment rate than the other agents. The agent without feasibility testing achieves sporadic success early on, but frequently "locks up" due to repeated attempts to duplicate infeasible mentor actions. The agent still manages to reach the goal from time to time as the stochastic actions do not permit the agent to become permanently stuck in this obstacle-free scenario. The control agent without any form of imitation demonstrates a significant delay in convergence relative to the imitation agents due to the lack of any form of guidance, but easily surpasses the agent without feasibility testing in the long run. The more gradual slope of the control agent is due to the higher variance in the control agent's discovery time for the optimal path, but both imitator and control converge to optimal solutions eventually. As shown by the comparison of the two imitation agents, feasibility testing is necessary to adapt implicit imitation to a heterogeneous actions context.

We developed feasibility testing and bridging primarily to deal with the problem of adapting to agents with heterogeneous actions. The same techniques, however, can be applied to agents with differences in their state space connectivity (these are equivalent notions ultimately). To test this we constructed a domain where all agents have the *same* NEWS action set; but we alter the environment of the learners by introducing obstacles that aren't present for the mentor. In Figure 6, the mentor's path is obstructed from the perspective of each learner. Movement toward an obstacle causes a learner to remain in its current state. In this sense, its action has a different effect than the mentor's action in this state.

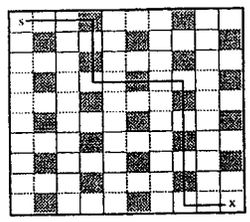

Figure 6: Obstacle Map and Mentor Path

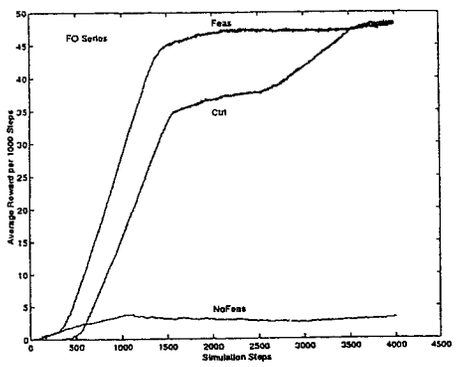

Figure 7: Interpolating Around Obstacles

In Figure 7 we see that the results are qualitatively similar to the previous experiment. Here, however, the top goal rate achieved by the observer with feasibility testing and the control agent is much higher because they are using the same action set as the mentor and can duplicate its path with out interpolating each action. The observer without feasibility has a more difficult time with this maze as the physical obstacles make it more difficult for the agent to achieve the goal purely by advancing due to the stochasticity of its actions. Essentially, however, local differences in state are well handled by feasibility testing.

Next we demonstrate how feasibility testing can completely generalize the mentor's trajectory. Here the mentor follows a path which is completely infeasible for the imitating agent. We fix the mentor's path for all runs and then we give the imitating agent a maze shown in Figure 8 in which all but two of the states the mentor visits are blocked by an obstacle. The imitating agent is able to use the mentor's trajectory for guidance and builds its own parallel trajectory which is completely disjoint from the mentor's.

The results in Figure 9 show that gain of the imitator with feasibility testing over the control agent diminish, but still marginally exist when the imitator is forced to generalize a completely infeasible mentor trajectory. The agent

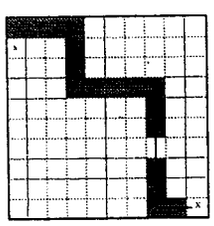

Figure 8: Parallel Generalization

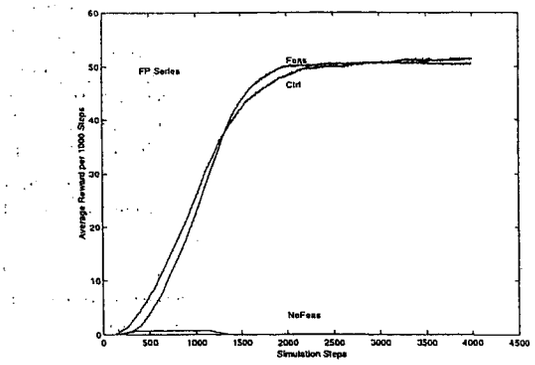

Figure 9: Parallel Generalization Results

without feasibility testing does very poorly, even when compared to the control agent. This is because it gets stuck around the doorway. The high value gradient backed up along the mentor's path becomes accessible to the agents at the doorway. The imitation agent with feasibility will conclude that it cannot proceed south from the doorway (into the wall) and it will then try a different strategy. The imitator without feasibility testing never explores far enough away from the doorway to setup an independent value gradient that will guide it to the goal. With a slower decay schedule for exploration, the imitator without feasibility testing would find the goal, but this would still reduce its performance below that of the imitator with feasibility testing. The imitator with feasibility testing makes use of its prior beliefs that it can follow the mentor to backup value perpendicular to the mentor's path. An aura of value thus clings to the mentor's path and the imitator can rapidly follow this aura to the doorway, make the necessary feasibility test at the doorway and then proceed to the goal.

As explained earlier, in simple problems there is a good chance that the informal effects of prior value leakage and stochastic exploration may form bridges before feasibility testing cuts off the value propagation that guides exploration. In more difficult problems where the agent spends a lot more time exploring, it will accumulate sufficient samples to conclude that the mentor's actions are infeasible long before the agent has constructed its own bridge. The imitator's performance would then drop down to that of an unaugmented reinforcement learner.

To demonstrate bridging, we devised a domain in which agents navigate from the upper-left corner, across a "river" to the bottom-right corner. The river runs vertically, is three steps wide and exacts a penalty of -0.2 per step. The goal state is worth +1.0. Without a long exploration phase, agents generally discover the negative states of the river and curtail exploration in this direction before actually making it across. If examine the value function estimate (after 1000 steps) of an imitator with feasibility testing but no repair capabilities, we see that, due to suppression by feasibility testing, the dark high-value states backed up from the goal terminate abruptly at an infeasible transition before making it across the river (see Figure 10). In fact, they are dominated by the lighter grey circles showing negative values. Once this barrier forms, only an agent with a very optimistic exploration policy will get to the goal,

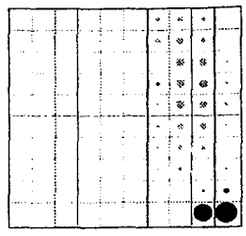

Figure 10: River Scenario

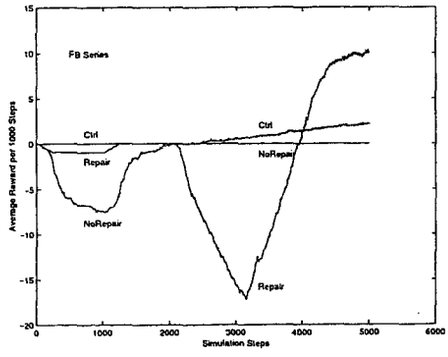

Figure 11: Utility of Bridging

and then only after considerable exploration. In this experiment, we apply a k-step repair agent to the problem with $k = 3$.

Examining the graph in Figure 11, we see that both the imitation agents experience an early negative dip as they are guided deep into the river by the mentor's influence. The agent without repair eventually decides the mentor's action is infeasible, and thereafter avoids the river (and the possibility of finding the goal). The imitator with repair also discovers the mentor's action to be infeasible, but does not immediately dispense with the mentor's guidance. It keeps exploring in the area of the mentor's trajectory using random walk, all the while accumulating a negative reward until it suddenly finds a bridge and rapidly converges on the optimal solution.[8] The control agent discovers the goal only once in the ten runs.

5 Conclusion

We have seen that feasibility testing extends implicit imitation in a principled manner to deal with the situations where the homogeneous actions assumption is invalid. Adding bridging capabilities preserves and extends the mentor's guidance in the presence of infeasible actions whether due to differences in action capabilities or local differences in state spaces. Our new approach makes use of a model to compute the actions an imitator should take without requiring that the observer duplicate the mentor's actions exactly. Our approach also relates to the idea of "following" in the sense that the imitator uses local search in its model to repair discontinuities in its augmented value function before acting in the world.

We see two major directions for future development of this line of research. The first is the application of this model to interesting some practical problems. We expect that combining our enhanced algorithm with more advanced exploration techniques and generalization capabilities will open up a broad range of tasks such as mobile robot navigation, process control, language learning and others. Another important direction involves extending our model to deal with partially-observable environments and to make explicit use of abstraction techniques.

[8] While repair steps take place in an area of negative reward in this scenario, this need not be the case. Repair doesn't *imply* short-term negative return.

References

Atkeson, C. G., & Schaal, S. (1997). Robot learning from demonstration. *ICML-97*.

Bakker, P., & Kuniyoshi, Y. (1996). Robot see, robot do : An overview of robot imitation. *AISB96 Workshop on Learning in Robots and Animals*, 3–11.

Billard, A., & Hayes, G. (1997). Learning to communicate through imitation in autonomous robots. *ICANN 97*, 763–68.

Dautenhahn, K. (1995). Getting to know each other – artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, *16*, 333–356.

Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics*, *7*(2), 247–54.

Hayes, GM; Demiris, J. (1994). Robotic learning by imitation. *The 3rd Inter. Conf. on Simulation of Adaptive Behavior, From Animals to Animats* UK.

Kaelbling, L. P. (1993). *Learning in embedded systems*. Cambridge: MIT Press.

Kuniyoshi, Y., Inaba, M., & Inoue, H. (1994). Learning by watching: Extracting reusable task knowledge from visual observation of human performance. *IEEE Tran. Robotics and Automation*, *10*(6), 799–822.

Mataric, M. J. (1998). Using communication to reduce locality in distributed multi-agent learning. *Journal Exp. and Theoretical Art. Intel.*, *10*(3), 357–369.

Mataric, M. J., Williamson, M., Demiris, J., & Mohan, A. (1998). Behaviour-based primitives for articulated control. *SAB-98, Zurich*, 165–170.

Mi, J., & Sampson, A. R. (1993). A comparison of the bonferroni and scheffé bounds. *Journal of Statistical Planning and Inference*, *36*, 101–105.

Mitchell, T. M., Mahadevan, S., & Steinberg, L. (1985). LEAP: A learning apprentice for VLSI design. *IJCAI-85*, 573–580.

Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, *13*(1), 103–30.

Nehaniv, C., & Dautenhahn, K. (1998). Mapping between dissimilar bodies: Affordances and the algebraic foundations of imitation. *EWRL-7* (pp. 64–72). Edinburgh.

Price, B., & Boutilier, C. (1999). Implicit imitation in multiagent reinforcement learning. *ICML '99* (pp. 325–34). Morgan Kaufmann Publishers, Inc.

Scheffe, H. (1959). *Analysis of variance*. New York: Wiley.

Seber, G. A. F. (1984). *Multivariate observations*. New York: Wiley.

Utgoff, P. E., & Clouse, J. A. (1991). Two kinds of training information for evaluation function learning. *AAAI-91*, 596–600.

Weisstein, E. W. (1996). *CRC concise encyclopedia of mathematics*. CRC Press. http://mathworld.wolfram.com/.

SOCIALLY COMPETENT BUSINESS AGENTS WITH ATTITUDE
Using Habitus-Field Theory to Design Agents with Social Competence[1]

Michael Schillo*; Steve Allen+; Klaus Fischer+, Christof T. Klein&

* Multi-agent systems group, Saarland University, Im Stadtwald, D-66123 Saarbrücken
+ Multi-agent systems group, DFKI, Stuhlsatzenhausweg, D-66123 Saarbrücken
& Department of Sociology, Saarland University, Im Stadtwald, D-66123 Saarbrücken
schillo@ags.uni-sb.de, allen@dfki.de, kuf@dfki.de, ctk@ags.uni-sb.de

Abstract

We will argue that social competence is an emergent mental phenomenon, and as such, there is no requirement to build discrete "social" modules into an agent. In fact, we argue that there are definite advantages to be gained from the emergent approach to social competence in complex, open, multi-agent environments. In order to capitalise on these advantages we need to design socially competent agents with the ability to reason on different levels (reactive, deliberative, meta) within complex social situations. By analysing the sociological theory of Pierre Bourdieu, we describe the design of a socially competent agent through the instantiation of a generic layered agent architecture. Our instantiation provides a methodology for specifying heuristics and parameters for different layers of such architectures. Furthermore, Bourdieu's habitus-field theory is hybrid in the sense that it tries to explain the effect of individual behaviour on societal structures and vice versa. This is where the great strength of the theory lies, and where we expect a useful cross-fertilisation of ideas into AI to occur. For as much as space permits, we will illustrate our argument with a scenario from the domain of shipping companies. This scenario is defined by its openness, diversity of agents as well as tasks and time restrictions. Our work leads us to the conclusion that building social agent architectures has definite engineering advantages, underlining the importance of this concept for both MAS and DAI research.

1 The need for socially competent agents in business

In this section we argue that there is a real need for social competence within multi-agent systems (agent-to-agent and/or agent-to-person) in complex business domains. We also argue that many of the challenges we face in such domains have direct relevance to the more general AI and computer science communities.

The TeleTruck CC project (Bürckert et al., 1998) at the German Research Centre for Artificial Intelligence (DFKI) addresses the problem of designing intelligent dispatch agents for shipping (road haulage) companies. Human dispatch agents not only map incoming orders to available trucks and drivers (the typical domain of centralised planning systems), but must also collaborate with the dispatch agents of other companies – to pass on orders that may be unprofitable for one reason or another, or simply not convenient. Collaboration between different companies creates an open, dynamic, and potentially hostile/competitive environment in which social competence plays a very important role. Our intelligent dispatch agents must not only deal with the social field of inter-agent collaboration, but must also address the social fields of agent-driver and agent-customer interaction. Each social field has its own logic, and its own set of resources (capitals) that may or may not be convertible – for example, a driver may be happy to give up a weekend for extra pay or a couple days holiday, but may resent doing so if it means missing his/her child's Birthday party.

A competent behaviour in our collaborative shipping scenario requires that a dispatch agent not only understands its own capabilities, but also the abilities, motivations, attitudes, goals, plans and the behaviours of the other competing/co-operating dispatch agents and truck or driver agents. For example, a dispatch agent needs to reason about how reliable the available drivers are, how beneficial business contacts to other dispatch agents are, whether it can trust other agents to fulfil the contracts they commit to, etc. Empirical research (interviews with human dispatchers) further shows that customer/driver models must also take into account the fact that certain customers may not want certain drivers to deliver their goods – adding yet another level of complexity.

[1] This work is supported by DFG (German National Science Foundation) under contract Fi 420/1-1.

Another model barely considered in current transport scheduling systems, is the model for drivers. Again, interviews with human dispatchers tell us that it is very important for the co-operation of the dispatcher with the truck drivers to take into account their personal preferences. Such preferences can be preferred routes, overnight stays, holidays, trucks, cargo etc. They again can put constraints on the total planning and the decisions as to whether to pass on orders to competitors or not. It is therefore important to know (a) when to consider driver preferences; and (b) when to override them. Also, many systems do not take into account the added interaction of driver, customer and vehicle. Certain orders can only be processed if the right driver and the right vehicle are available at the right place – this is especially true when shipping companies deal with transporting food, highly explosive liquids or containers and only have a limited number of vehicles which are available to transport these special kinds of cargo.

Not only does a dispatch agent need to meet the constraints in its planning activities, it also needs to know how important a particular constraint is for medium- or long-term goals. For example, if it discovers that there is no solution for the current set of constraints, it needs to know which constraints can be relaxed. Decisions therefore require an understanding of the relative importance of qualitatively different constraints, which in turn requires an understanding of the relative convertibility of resources between the different social fields.

The complexity of the shipping domain is such that the *real* plan space of a dispatch agent is far greater than that covered by existing route planning and cost minimisation dispatch systems. This is clearly reflected by the fact that all the existing systems on the market require human operators to provide the missing levels of social competence. What shipping companies need are socially competent agents that can act autonomously.

Many of the requirements of socially intelligent agents are strikingly similar to the requirements of autonomous agency: agents hold inconsistent beliefs; have multiple competing concerns which are qualitatively diverse; and must be robust in the face of hostile and unknowable environments. In building socially competent intelligent agents, we will inevitable address many of the same problems faced by researchers in the field of intelligent autonomous agents – we hope that our approach will provide insights that are beneficial to both communities. In this sense, we also believe that there will be significant synergy between research on socially situated agents and research on *bounded rationality*, or *bounded optimality* (Russell 1997) in the more general AI and computer science communities.

2 What is social competence?

Dautenhahn and Edmonds (1998) argue that the intelligence of a socially situated individual, and social interaction, are inseparably intertwined. They make a strong case for a bottom-up approach to modelling socially intelligent behaviour, which involves working out the principal processes from which higher order social competence will follow. Our work fits within this framework.

In order to avoid the somewhat difficult to define concept of social intelligence, we will use the term *social competence*. Research on *computational organisational theory* (e.g. Carley and Gasser, 1999) tells us that the main reason for the dynamics of complex behaviour within large organisations lies in the unequal distribution of resources between agents. Which can also be stated as "if every agent had all the resources it needed, there would be no complex societal interaction." The striving of each individual to get hold of the needed resources, the communication, negotiation and action that is necessary to gain access to goal-satisfying resources will create the complexity of the behaviour of the organisation as a whole.

We consider agent behaviour to be socially competent, if it manages to recognise the strategies which lead to access to resources – this is the cornerstone of social competence. More competence is necessary if the resources become more complicated, i.e. are made up of different types, and/or there exists different modes of exchange between certain resources etc. Another difficulty in recognising which resources need to be accessed, comes from the need to know the connection between goals and resources. Furthermore, the agent must recognise that its environment is made up of a number of agents that must be dealt with individually, and cannot be treated as one homogeneous entity.

One last issue for social competence we would like to raise at this point, is the fact that every socially competent agent must deal with one outstanding kind of resource, namely time – which brings us back to the issue of bounded rationality discussed in the last section.

To summarise: the dynamic properties of societies rely on (a) the access to required resources and (b) the competence to deal with the acquisition of non-accessible resources. In this sense, agent interaction is not only the exchange of information, but also an instrument "to influence others, change their goal-balance, and induce them to adopt one's goals" (Castelfranchi and Conte, 1996). Interaction and the dependence on others means that agents must model explicitly the effects of themselves and others on society and take these effects into account when considering long-term plans. We therefore need a methodology that provides us with concepts to capture what "social" means, and to enable us to analyse and design socially competent

agents. A theory which provides such a methodology is the theory of Pierre Bourdieu.

3 How do humans achieve social competence: the theory of Pierre Bourdieu

Pierre Bourdieu' work emphasises not only the structural aspect of society, as represented in his model of social space, but also the action aspect of social life. His theory is known as the theory of habitus and field, which is intended to overcome the "clash" of social theory in the micro and macro domains. In his view, this clash results from sociologists "creating" antonyms by using either "objectively" formed structure (constitutional elements of society as a system, e.g. Luhmann, 1995), or "subjectively" formed actions (constitution of social life by interacting e.g. Berger and Luckmann), in their explanations. In breaking with these antonyms, the approach of habitus and field develops an interdependence of structure and action – instead the prevailing exclusive or treatment.

3.1 Sociocentrics of cognitive structures

The basic assumptions of habitus-field theory are derived from a structural analysis of "primitive" societies conducted by Durkheim and Mauss. Observing primitive societies and their structures, Durkheim and Mauss reported a coincidence of objective social structure and cognitive structures of individuals. According to their findings, societies that are not able to give a solid mechanism which socially determines the systems of classification, will fail when attempting a shift to an advanced society. This thesis is known as the sociocentrics of cognitive systems. The existing cognitive systems are deduced from the global social system, with the categories of reason underlying the collective ideas built according to structures of the social group. Up to this point, Bourdieu's theory agrees with the thesis of Durkheim. However, Bourdieu was able to extend the habitus as a hybrid dialectic concept, linking both societal and cognitive structures (see next section). The habitus is theoretically used to explain the coincidence of social and cognitive structure by sociogenetics – i.e. transfer of group-specific shared schemas through language and the educational system (which depend to some extend on genetics).

At the heart of the sociocentrics of cognitive systems lies the basic assumption that the formation of classifications (as categories of perception) – which are based on structures of a group-specific segmented social world – organise and regulate actions in social practice (Bourdieu and Wacquant, 1992: pp. 30-34).

3.2 Action Theory: theory of practice and habitus

This practice is according to Bourdieu the product of the dialectic relation between a situation and a system of lasting and transferable dispositions of an social actor, called habitus. With his habitus concept, Bourdieu tries to capture a system of dispositions "that by integrating all former experiences seems as a matrix of action, perception and reasoning and is based on the analogous transfer of schemes (...) and allows to fulfil infinite differentiated tasks" (Bourdieu , 1977).

The action theory which Bourdieu proposes (including the habitus), argues that the underlying reason for most human action lies far away from what we know as intentional rationality – Bourdieu identifies the motives of actors as acquired dispositions. The variation of (individuals) habitus derive from the objective societal structure, primarily depending on, and mediated by, the group heritage. Thus, similar conditions of existence, which could be described objectively by group-specific positions in social structure, lead to quiet similar habitus (group-specific dispositions). These constitutive social differences will be incorporated (internalised) by processes of socialisation and enculturation using the human body. The main role of the body lies in the support of memory for those internalised, and position variable, collective schemes. The schemes themselves were sociohistorical grown, as well as the co-responding social structure. The transfer of group-specific shared schemes is mainly carried out by language and the education system – creating the coincidence of objective societal structures with the cognitive structures of individuals. Social differences and mental schemes are homologous because they are co-related genetically.

In other words: By their habitus, actors are themselves the owners of basic symbolic systems of classification of their society. The dispositions – a homologous representation of a social-structured space the actors move in – allow each actor to act as if she/he knows what is to do in almost every situation. The cognitive system and the social system form a perfect synergy. The social system allows an actor to act appropriately in his/her existing societal environment, with the criteria for success proofed by other actors – i.e. by interactive processes (see also field). The shared collective social constitution then allows any small differences to be perceived as the natural properties of individuals (i.e. gender), and in this way can be taken for granted. By transferring the basic cognition schemas to other situations or contexts, actors are then able to act appropriately in new situations – on the basis of the dispositions internalised from their existing societal environment.

3.3 Habitus concept: incorporated dispositions for perception and action

One part of the concept habitus allows us to explain the reproduction of a concrete culture and its differences. In this respect, Bourdieu uses the term structured structure for the habitus built by structure. Referring only to this dimension of the concept it seems that habitus is nothing more than that determined by objective socio-cultural structure. In this structural respect, the habitus is caused by the embedded cultural and could not be changed in a fundamental manner (so as to say we are determined by a structured habitus, conditioned by the existing cultural values and forms).

However, viewing this as the only interpretation of habitus, is a misinterpretation (Bourdieu and Wacquant, 1992: p. 19). Therefore it is important to point out the gains Bourdieu make by developing antonyms: The habitus is the basic concept for constructing a theory of structures, which is able – in contrast to most other structural theories – to answer the question of how acting may escape the structural pressure (Lemert, 1990: p. 299, op. cit. Bourdieu and Wacquant, 1992: p. 166). This is the second aspect of habitus: the aspect of structuring social structure by actors through, and according to, their specific dispositions.

The habitus concept can therefore be divided in to two aspects: One depends on the incorporation of the existing historically grown structure. In so far, behaviour of actors seems to be determined by the process of internalisation objective cultural patterns (see pattern variables and his normative paradigm in Parsons, 1964). The other aspect is that if an actor has internalised the structures of social life, he may interpret them, and by the act of reproducing them, change them (see A. Schütz, 1940, and the interpretative paradigm).

He [the habitus] *is a socialised body, a structured body, a body, which has internalised the immanent structures of a world or of a specific sector of this world, of a field, structuring same actors' perceptions of and actions in this* Bourdieu, 1998: Practical reason: On Theory of Action).

3.4 The field as a social context

Habitus and field are effectively related (Bourdieu and Wacquant, 1992: pp. 34-49). Bourdieu develops the conditions of "objective" structure in his field concept as a model of social space. According to several sociologists[2], the process of differentiation in modern societies is continuing, and Bourdieu represents this ongoing process in his model of social topology. The task of this model was to re-construct the social space in its actual forms and differences. The basic characteristic of this space can be found in the reciprocal relations of the objects which are included. Social sciences in Bourdieu's perspective try to objectify the main principles of differentiations according to observed social differences. Thus, an investigator is able to explain the statistical distributions in a given societal structure.

The social world may conceptualised as a multidimensional space, which empirically is constructed by the differentiation, by which the given societal universe could be explained, or in other words, throughout the discovery of the forces or forms of capitals, which like jokers in a game of cards, becoming in this specific universe effective or could, i.e. in that fight (or the competition) around rarely goods, which are located in this universe (Bourdieu and Wacquant 1992: p.106).

The distribution of social energy – as Bourdieu also names the downward explicated capitals – corresponds to the distribution of attributes. By these processes, their owners gain force, power, and on the basis of both, profit.

3.5 Autonomy of fields and interests

In Bourdieu's view, the whole social space is partitioned into several universes, which he calls "fields". Each social field gains (produces and reproduces) its own identity by actors interested in objects which are part of the game in this special field. Each field (e.g. the field of economics, art, science etc.) produces its own "nomos" (Greek: rule), which is independent from all other fields – i.e. each field is autonomous. This autonomy is created by competition of the interested actors, investing in the specific field and its objects. E.g., if each artist produces her/his objects and actions for only explicit economic reasons, there would be no longer any social difference between his/her art and the actions of an undertaker or workman who are producing goods for the demands of a market. If this were really the case, then the field of art could be defined according to the tautological rule of the economic field (business is business): "art is business". But, by stressing the basic difference between the two fields in a mostly unconscious way (see: habitus), each actor engaged in the countless and various objects of art (artists, Galleries, "gourmets" of art) creates the field and restores its social autonomy.

This process of differentiation and becoming autonomous, leads to the rise of universes which have distinguished, not reducible basic rules [...] and so they are decision fields [battle-fields] *for special forms of interests. What concerns people in the field of science and leads them to competition is quiet different to its equivalent in the economic field* (Bourdieu, 1977). E.g. "investments" of "social capital" in the "social sub-field" (a sort of family through parental care of actors).

[2] See Durkheim, Weber, Simmel and for example Luhmann's conception of the social system (functional differentiation) as divided into functional subsystems, organisational systems and interactional systems

The actors being involved in a field and its objects, do not need to intentionally plan their goals of action (e.g. game-theory), nor do they need act exclusively for economic benefits (see fields with interests on the non economic interests, e.g. religious field). The future is something they anticipated in their presence by the help of their dispositional practical sense. The shape of interests depends on the objects in the game.

Social actors who possess a sense for the game and have incorporated the countless practical schemas of perception and evaluation, which work as instruments of construction of reality, as principles of observation and tidiness of the world in which they move, do not need to put (...) the aims of their practice as a purpose (Bourdieu, 1979: p. 144).

The better the internalised schemas (the dispositions of an individual) fit with the habitus (the structure of a specific field mediated by the "community" of engaged actors – i.e. the scientific community of CS), the greater the chance of an individual to become a "master" in that field. The group-specific generating process of individual habitus explains both the processes of social integration (e.g. forms of co-operation) and the processes of differentiation (e.g. forms of conflict).

3.6 Practical logic of the field: capitals as different forms of social resources

According to Bourdieus investigations[3], the forces of social life cannot be reduced to s single dimension. The models of utilitarian approaches are in his view much too simplistic to explain social phenomena. With the logic of the practice, Bourdieu broadens the narrow model of the utilitaristic perspective (comp. Bourdieu/Wacquant, 1992; p. 147). In an actual social universe (e.g. French society) the only accepted form of resources by utilitarian approaches (the economic capital), will not suffice. Thus Bourdieu adds other forms of societal energy: cultural, social, and symbolic capital. The actors are then distributed over the whole space to variable degrees based on the possession of convertible capitals (includes possibility of loss same as extensions of ones stock).

3.7 Three Dimension of Social Positions

Bourdieu differs three dimensions of distribution: (1) according to the whole volume on capitals one possess; (2) according to the composition of their stocks of capitals (especially the relation of economic and cultural capital); and (3) according to the development of their whole capital in time, i.e. according to their career in social space (Bourdieu, 1997: pp.108/109).

Competition in a field leads to a permanent and latent conflict of actors for transferring one form of capital into another. The substrate of the economical capital is money – it is objectified as "possession", and institutionalised in the form of ownership rights. The probability of converting money into other forms is "high", with risks of deterioration lying in social crisis (wars, revolutions, economical crisis).

A similar analysis can be made for the other capitals, as for example with the capital of culture (or more precisely, the capital of information). The substrate of cultural capital is "knowledge". Cultural goods and knowledge are its objectified forms, and it is institutionalised by titles of education. The substrate of social capital can be identified as relationships, which in its objective form, Bourdieu terms this capital "networks". Its institutionalised forms are titles of aristocracy as individual predicates and the status of profession as an collective schemas. To convert social capital has little reliability and is risky, but often necessary.

4 How does Bourdieu's theory match this design and help to build socially competent agents?

Bourdieu's theory describes in a natural way how societies evolve and adapt. These processes unfold with the need of the individual to adapt his/her habitus to the logic of the field, in order to pursue interests and to benefit from previous investments. Therefore, using the sociological theory of Bourdieu will automatically lead to an artificial society that cannot deny its anthropological origin. This again is a strength of using the concept of habitus to design agents – providing a framework that allows us to explain the behaviour of a given system and fine-tune its design more accurately than without the conceptual apparatus of this human adequate theory.

If we briefly look at our scenario again, we can identify four different fields of interaction for the dispatch agent. The field of interaction with: (a) customers; (b) drivers; (c) trucks; and (d) the dispatch agents of competing companies. When interacting in these fields, the dispatch agent needs to take into account their relative importance, preferences, reliability and persistence in relation to their commitments. This interaction depends on the field. For example, the dispatch agent will pass on or receive orders from other dispatchers, give orders and receive information about costs from the trucks. Finally, the dispatch agent will be informed about preferences by the drivers and will give them orders in the form of route plans. The dispatch agent's habitus will be shaped by what knowledge has been implemented off-line and the experiences that it has made during runtime in a variety of different situations. To take up again the second part of Bourdieu's habitus, we can say that the agent's habitus will also have an effect on the structure of the whole society of agents. For example, if the percentage of agents which

[3] For an extensive overview of Bourdieu's works see Bourdieu and Wacquant 1992; p. 295-307

are neither reliable nor co-operative, is too large, we will get a society of agents with a different kind of social structure than if the distribution of resources would rule out conflicts about resources.

The capitals in this setting include the social capital, i.e. the relationships ("contacts") to certain dispatchers that have proven to be of mutually beneficial or drivers that have shown to be trustworthy, punctual etc. There is also the capital of economics which mirrors (for example) the amount of trucks available or the amount of money that has been gathered in the past. An example for the information capital would be the knowledge about the preferences a customer has acquired during previous interactions and can be exploited for specifically tailored future services. Conversion between these capitals is manifold. Contacts can be used to find out about customer preferences before making an offer for a certain order, or knowing the preferences of a customer can be used to decide which driver must be sent to him/her. Social capital in the form of "owes me a favour" can be converted to make a driver accept an order, which he would otherwise have rejected (weekends etc.). Of course there is also the traditional conversion between capitals using economic money, like buying information, stabilising relationships to customers by reducing prices etc. This list of capitals and their conversion is not exhaustive.

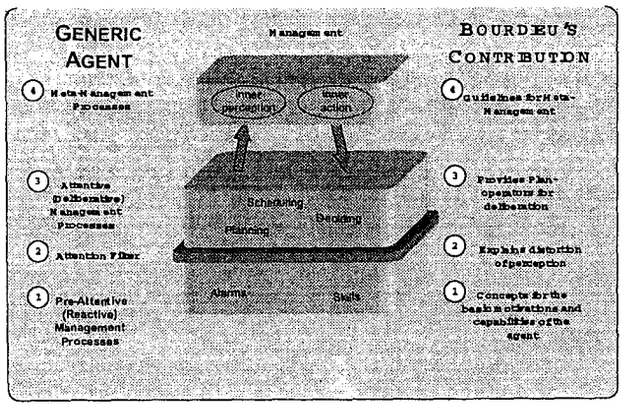

Figure 1 The contribution of Bourdieu's theory to social agents.

Having outlined the application of Bourdieu's theory to the fields in the shipping scenario, we will now apply it to a model of a generic multi-layer agent architecture – which we believe fits to many modern agent architectures (Allen, to appear; Müller, 1996; Jung 1999; Sloman, 1999; etc.). This generic architecture consists of three layers of reasoning: a reactive layer; a deliberative layer; and a meta-deliberative layer (see Figure 1). The pre-attentive/reactive layer and the deliberative layer are separated by a filter to stop excessive interruption of deliberative/attentive layer by insignificant events in the environment. An example for a methodology for such a filter could be Sloman's Attention Filter Penetration Theory (Sloman, 1992). The "meta" layer deals with the management of the deliberative layer (monitoring and control of the deliberation). In some architectures the arrangement is slightly different, however these "processing levels" are in one way or the other represented.

We will now show the difficulties of breaking down the aforementioned abstract parts of the architecture into implementable concrete algorithms. Furthermore, we will show how habitus-field theory helps us to conceive concepts and determine values that can be calculated or learnt respectively during runtime.

Let us first deal with the lowest level of the architecture. Bourdieu would label this level as the level in which the basic interests/concerns of the agent in its environment are realised. Some interests in the field (i.e. increase blood sugar level if hungry) provide goals and therefore start the process of trying to get access to certain resources. In Bourdieu's terms this would be the motivation to take part in the game of this social context. Other interests that may be hindering the cognitive process are specified by the inability of the individual to cope with the current situation, as the habitus (the set of action dispositions) can neither be changed rapidly nor is the habitus of this level completely known to the individuals and escapes conscious manipulation.

Every reactively generated goal (or processed percept) that is to be recognised by deliberative processes, must pass through the attention filter. We believe that the attention filter is a very good concept to describe the fact that certain aspects of a situation are perceived by some individuals and neglected by others (for a discussion see Köstler, 1968). In practice, it turns out to be rather difficult to define the insistence parameters for the different percepts that in the end decide which percept will receive any attention. Bourdieu's theory gives more insight into the heuristics that need to be chosen for defining the parameters. From his point of view it is evident that the attention that is paid to perception is restricted by the experiences (acquired habitus) and the education or culture (incorporated habitus) of the individual. The primary/reactive layer's purpose is to provide good candidates for consideration by the deliberative layer above. This is especially useful when evaluating contingent situations, i.e. situations where a number of different possible worlds need to be considered simultaneously. Which candidate for consideration is chosen in the end depends on the individuals habitus, its incorporated dispositions for action and perception.

Once a percept has received attention, it is dealt with by the deliberative layer of the architecture, i.e. we talk about the layer that is able to reason. Like Conte (1997), we argue for a heterogeneity of rationality. In her words, we need to introduce a substantial differentiation, namely a qualitative heterogeneity among individual agent's goals. To do this, a goal-based rather than preference-based view of endogenous motivations should be chosen (this fits well with Bourdieu's idea of broadening). The difference between goals and prefer-

ences is fundamental: the essential difference between them is on the qualitative vs. quantitative characterisation. While preferences are quantitatively defined, goals are symbolic and qualitative notions. Unlike the former, they allow agents to be heterogeneous. Only this choice can guarantee that a variety of social actions can be described and predicted – in terms of rational decision theory applied to social settings, only defection or co-operation is possible. Social life is interspersed with different types of pro-social action, from influencing to exchange to co-operation. Conte argues that utility cannot actually account for such a variety, while a qualitative notion can.

So when we talk about symbolic representation of the goals of an agent in its deliberative layer, we also need to talk about some kind of calculus that is able to manipulate these symbols. Here a suitable AI technique would be automated planning. Crucial for planning is the provision of plan operators (we would like to stress this point, as the division of actions into plan operators is not trivial). If the system is not aware of the available and necessary operators, no algorithm will be able to find a satisfying plan. This is crucial to the success of the agent in an open system and places an emphasis on the ability to recognise plan operators. If we do not expect agents to learn *everything* from scratch, we need some description that underlies plan operators. Again, this is were Bourdieu comes in. He lays out that every field (i.e. social context) has its own number of capitals (or, in AI terms, resources).

The goals of an actor can be represented in a description of which capitals the actor wants to increase. Following Bourdieu, the description of the plan operators is merely the description of how one capital can be converted into another. For example, buying a prestigious car is the conversion of some economic capital (money) into another form of economic capital (possession) and symbolic capital (a car that stands for status). Thus the problem can be reduced to that of specifying the number of capitals and the conversion matrix. With this concept at hand, it is far easier to socially learn the plan operators, either by imitation or by advice taking. Of course, trial and error is still an option. However, knowing what has to be learned in principle (namely the capitals and their convertibility) will make the task a great deal more feasible. Yet another way to enable the agent to cope with the complex social situation is to build in the most important parameters. Bourdieu also captures this kind of procedure. He calls it the incorporated part of habitus, with all the problems that come with it, i.e. the problem of changing a maladjusted part of the habitus in contrast to the desired continuity of the habitus.

The top level of our architecture deals with the management of the deliberative layer. This is the level on which the agent monitors its own deliberation processes and tries to work out which of them lead to blind alleys, are ineffective, or deadlocked. This layer is about controlling the thought process. As well as with the other layers, the processes on this level should be made flexible enough to be able to react to changes in the environment and use plan operators as they become available (e.g. by changes in the field) etc. Therefore, just as in the other layers, the processes must exhibit some features of habitus, they are a result of their own history.

5 Conclusions

The theory of Bourdieu helps us in instantiating a generic agent architecture. We argued with Bourdieu's theory, that the history of the individual and what it experiences results in dispositions to certain actions, ways of perception and considerations. These dispositions may be incorporated or imitated, i.e. learned by observation and acquired by advice. We deny that there is a habitus module somewhere in a social agent architecture, instead, we argue that the habitus is the result of processes that adapt to the environment on all layers of the generic architecture. This emphasises the importance, and the influence, of the culture of the agent society on the behaviour of the individual. When looking from the other direction, Bourdieu's habitus-field theory predicts how interaction in a society will change as compatible and incompatible habitus are forced to interact by their interest in the social field.

Acknowledgements

We would like to thank Sociologists Michael Florian, Andrea Dederichs and Frank Hillebrandt from the Department of Technology Assessment at the Technical University of Hamburg-Harburg for most fruitful and enlightening discussions.

6 References

Allen, S. (To appear). *Concern Processing in Autonomous Agents*. PhD Thesis. School of Computer Science and Cognitive Science Research Centre, University of Birmingham.

Bürckert, H.-J., Fischer, K. and Vierke, G. (1998). Transportation Scheduling with Holonic MAS -- The TeleTruck Approach. *Proceedings of the Third International Conference on Practical Applications of Intelligent Agents and Multiagent systems (PAAM'98)*.

Berger, P. and Luckmann, T. (1966). *Social Constitution of Reality*, Anchor.

Bourdieu, P. (1977). *Outline of a Theory of Practice*, Cambridge University Press.

Bourdieu, P. (1997). *Der Tote packt den Lebenden*. VSA Hamburg.

Bourdieu, P. and Wacquant, L. (1992) *Invitation to Reflexive Sociology*, University of Chicago Press.

Carley, K. and Gasser, L. (1999). Computational Organizational Theory. In Gerhard Weiß, *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence.* MIT Press.

Castelfranchi, C. and Conte, R. (1996). Distributed Artificial Intelligence and Social Science: Critical Issues. In O'Hare, G.M.P. and Jennings, N. R. (eds.) *Foundations of Distributed Artificial Intelligence*, New York, p. 527-542.

Conte, R. (1997). Diversity in rationality. A multi-agent perspective. In Gilbert, N., Müller, U., Suleiman, R. and Troitzsch, K. *Social Science Microsimulation: Tools for Modelling, Parameter Optimization, and Sensitivity Analysis.* IBFI, Dagstuhl-Seminar-Report 177.

Edmonds, B. and Dautenhahn, K. (1998). The Contribution of Society to the Construction of Individual Intelligence. In Edmonds, B. and Dautenhahn, K. (eds.), *Socially Situated Intelligence: a workshop held at SAB'98*, August 1998, Zürich. University of Zürich Technical Report, 42-60.

Jenkins, R. (1993). Review of 'The Logic of Practice', Man 28(3), 617-18.

Jung, C. G. (1999). *Emergent Mental Attitudes in Layered Agents.* Intelligent Agents V, volume 1555 in LNAI, Springer.

Köstler, A. (1968). *The ghost in the machine.* Macmillan, New York.

Lemert, C. S. (1990). The Habits of Intellectuals: Response to Ringer; in: *Theory and Society* 19, no.3, p. 295-310.

Luhmann, N. (1995). Social Systems, Stanford University Press.

Müller, J. (1996). *The Design of Intelligent Agents: A Layered Approach.* Lecture Notes in Artificial Intelligence 1177, Springer Verlag.

Parsons,T. (1964). *The social system.* Routledge & Kegan Paul, London.

Russell, S. (1997) *Rationality and Intelligence.* Artificial Intelligence, 94, p. 57-77.

Sloman, A. (1992). Proloegomena to a Theory of Communication and Affect. In Orthony, A., Slack, J. and Stock, O. (Eds.) *Communiation from an Artificial Intelligence Perspective.* Springer, Heidelberg.

Sloman, A. (1999). Architectural Requirements for Human-like Agents Both Natural and Artificial. (What sorts of machines can love?). In K. Dautenhahn (Ed.) *Human Cognition And Social Agent Technology*, John Benjamins Publishing.

Schütz, A. (1940). *Phenomenology and the social sciences.* Cambridge.

Introducing Emotions into the Computational Study of Social Norms

Alexander Staller and Paolo Petta
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna, Austria (EU)
{alexs, paolo}@ai.univie.ac.at

Abstract

We argue that modelling emotions among agents in artificial societies will further the computational study of social norms. The appraisal theory of emotions is presented as theoretical underpinning of Jon Elster's view that social norms are sustained not only by material sanctions but also by emotions such as shame and contempt. Appraisal theory suggests the following twofold relationship between social norms and emotions: First, social norms play an important role in the generation of emotions; second, emotion regulation depends heavily on the influence of social norms. Based on these insights, we present an emotion-based view on the influential study by Conte and Castelfranchi (1995); without mentioning emotions, they argue that a function of social norms is aggression control. Appraisal theory offers a principled framework for the development of TABASCO, a three-layer agent architecture incorporating social norms. At the macro level, the computational study of social norms can profit by economic and sociobiological theories, which suggest that emotions play an important role in sustaining norms of cooperation and reciprocity. We show how appraisal theory can serve as a link between the macro and micro levels, and summarize the potential benefits from the development of TABASCO.

1 Introduction

Imagine you are invited to dinner. You think this will be an informal event and put on your jeans. However, you soon realize that you are the only guest who is not wearing a dinner jacket or an evening dress. The other guests look contemptuous and avoid talking to you. You feel the tendency to hide, which is a sign of being ashamed.

This example suggests that the violation of a social norm can trigger emotions such as contempt and shame. In this paper, we will elaborate on the relation between social norms and emotions and argue that the computational study of social norms can profit by modelling emotions among agents in artificial societies. In section 2 we will present an emotion-based definition of social norms by Elster (1989). Section 3 is devoted to the appraisal theory of emotions, suggesting that social norms play in important role both in emotion generation and emotion regulation. Appraisal theory provides us with the theoretical underpinning to present an emotion-based view of the study by Conte and Castelfranchi (1995) in section 4. In section 5 we will outline TABASCO, our appraisal-based agent architecture, and present theoretical considerations for incorporating social norms into TABASCO. Section 6 contains a review of economic and sociobiological theories suggesting that emotions play an important role in sustaining norms of cooperation and reciprocity. In section 7 we suggest that appraisal theory can serve as a link between the macro and micro levels. Section 8 concludes the paper by summarizing how the computational study of social norms could benefit from the development of the TABASCO architecture.

2 An Emotion-Based Definition of Social Norms

The example in the introduction suggests that emotions such as contempt and shame play an important role in sustaining social norms. Elster (1996, 1999) has taken this view. He defines social norms as injunctions to behaviour with the following features:

First, social norms are *not outcome-oriented*. In the simplest case they are of the type 'Do X' or 'Do not do X'. If the imperative expressed by a social norm is conditional, then it is not future-oriented. For example it is of the type 'If others do Y, then do X'. By contrast, rational action is concerned with outcomes. A rational, self-interested actor follows the maxim 'If you want to achieve Y, do X'.

Second, for norms to be *social*, they must be shared by other people. Some norms are shared by all members of the society, while other norms are more group-specific. Another respect in which norms are social is that other people are important for enforcing them through sanctions.

Third, social norms are not only sustained by the sanctions of others, but also by *emotions*. The violation of a social norm can trigger negative emotions such as shame or guilt in the norm violator, even if nobody can observe the norm violation. So emotions arise as negative internal consequences of a norm violation and thus sustain social norms in addition to external sanctions.

On this account, emotions do not seem to be a necessary part of a system of social norms. The enforcement

of social norms appears to be overdetermined by sanctions and emotions. But Elster (1996, 1999) argues that emotions are crucial for the operation of sanctions. A person who is imposing sanctions on the norm violator is driven by emotions such as contempt or disgust. A sanction may be just a subtle expression of such an emotion, e.g. a facial expression. Even if the norm violator does not suffer any material loss, the sanction is still effective because the norm violator "will see the sanction as a vehicle for the emotions of contempt or disgust and suffer shame as a result" (Elster , 1999, p. 146). The introductory example is a case in point.

Elster's view presupposes that social norms play an important role in the *generation* of emotions such as contempt and shame. In addition, he notes that emotions and their expression may be *regulated* by social norms. As an example he puts forward the norm against laughing at funerals (Elster , 1996).

Is there any theoretical support for this twofold relation between social norms and emotions? Indeed, appraisal theory – especially Frijda's (1986) approach – explicitly deals with the role of social norms in the generation and regulation of emotions. In the next section, we describe appraisal theory in more detail.

3 The Appraisal Theory of Emotions

After having long been dismissed as irrational and of no utility, emotions are now seen as a key element in successful coping with a non-deterministic, dynamic, and social environment. Appraisal theory emphasizes that this coping depends on the continuous monitoring of the relationship between the individual and the environment. Its central tenet "is the claim that emotions are elicited and differentiated on the basis of a person's subjective evaluation or appraisal of the personal significance of a situation, object, or event on a number of dimensions or criteria" (Scherer , 1999, p. 637). Thus, appraisal theory explains why the same event can give rise to different emotions in different individuals, or even in one and the same individual at different times. Conversely, appraisal theory offers a framework for the identification of the conditions for the elicitation of different emotions, as well as for understanding what differentiates emotions from each other.

3.1 Appraisal Criteria

Many theorists have been trying to specify the criteria according to which a situation is appraised (Roseman, 1984; Scherer, 1984; Smith and Ellsworth, 1985; Frijda, 1986; Ortony et al., 1988; Lazarus, 1991). There is a high degree of consensus with respect to these criteria. According to van Reekum and Scherer (1997, pp. 259-260), these include "the perception of a change in the environment that captures the subject's attention (novelty and expectancy), the perceived pleasantness or unpleasantness of the stimulus or event (valence), the importance of the stimulus or event to one's goals or concerns (relevance and goal conduciveness or motive consistency), the notion of who or what caused the event (agency or responsibility), the estimated ability to deal with the event and its consequences (perceived control, power or coping potential), and the evaluation of one's own actions in relation to moral standards or social norms (legitimacy), and one's self-ideal."

The postulate of appraisal theory is that specific profiles of appraisal outcomes on these criteria determine the nature of the ensuing emotion. Scherer (1999, p. 639) provides a table of theoretically contended appraisal profiles for anger/rage, fear/panic, and sadness.

3.2 The Appraisal Process

The description of the appraisal criteria in abstract, conceptual terms, often represented as a series of questions to be evaluated, led many critics to assume that the appraisal process is necessarily deliberate and conscious. For example, Zajonc (1980) criticized the "exaggerated cognitivism" of appraisal theory. In response to this criticism appraisal theorists pointed out that the appraisal process largely occurs nonconsciously and involves perceptual processing. The fear of a tiger jumping out of the bush is certainly not elicited by a conscious evaluation of appraisal criteria, but by fast perceptual processes. The appraisal process involves perceiving the "affordance" (Gibson, 1979) of stimulus events for one's coping activities (Smith and Lazarus , 1990; Frijda , 1993).

An affordance is defined by Gibson (1979, p. 127) as "what it offers the animal, what it provides or furnishes, either for good or ill." The general idea is that an animal actively perceives meaning in the environment without further interpretative cognitive processing. So there is a direct coupling between perception and action. McArthur and Baron (1983) apply the affordance concept to social perception, e.g. to emotion perception, impression formation, and causal attribution.

Leventhal and Scherer (1987) include perceptual processing in their model of the appraisal process. They suggest a hierarchical processing system consisting of three levels: sensory-motor, schematic, and conceptual. The sensory-motor level is based on innate hard-wired feature detectors which can give rise to emotional reaction directly. The schematic level is based on schema matching. The conceptual level involves reasoning and inference processes that are abstract, active, and reflective.

Building on the model by Leventhal and Scherer (1987), Smith and Kirby (2000) suggest a model of the appraisal process in which perceptual processing is complemented by associative processing (i.e., schematic processing) and reasoning (i.e., conceptual processing). Associative processing is a fast, automatic, parallel, and me-

mory-based mode of processing. As memories of previous experiences are activated, appraisal meanings associated with them are activated automatically. In contrast, reasoning is a relatively slow, controlled, and serial process that actively constructs appraisal outcomes. A novel feature of this model is the existence of so-called *appraisal detectors*. They monitor appraisal information generated through associative processing and reasoning, in addition to perceptual information, and generate an emotional reaction. The appraisal detectors are assumed to model the function of the amygdala, which plays an important role in the elicitation of fear (LeDoux , 1996) and presumably of other emotions as well.

The view of appraisal as a multi-level process corresponds to the recent trend towards multi-level theories of cognition-emotion relations in the areas of clinical psychology, neuropsychology, and the study of memory (Teasdale , 1999). Van Reekum and Scherer (1997) discuss the pertinence of such theories for a model of the appraisal process in more detail.

3.3 The Emotion Process

Throughout the rest of section 3 we follow Frijda (1986), a main proponent of appraisal theory. All citations refer to Frijda (1986).

Appraisal is the first step of the emotion process. For successful coping with the environment the appraisal outcome must have an effect on the actions of the individual. But appraisal does not lead to action directly. Instead, appraisal is followed by an impulse, i.e., the instigation of an action tendency. Action tendencies "are states of readiness to achieve or maintain a given kind of relationship with the environment. They can be conceived of as plans or programs to achieve such ends, which are put in a state of readiness" (p. 75). For example, the action tendency of fear is avoidance. An impulse involves shifts in control over behaviour, attention, and resources that are referred to as the "control precedence" feature of emotion. Frijda et al. (1989) have established significant relations between particular appraisal patterns and action tendencies. Thus emotions can be regarded both as experiences of forms of appraisal and as states of action readiness. The final step of the emotion process is the generation of cognitive or overt action, possibly in the form of mostly expressive behaviour such as facial expressions.

Frijda emphasizes the importance of emotion regulation. All steps of the emotion process sketched so far are subject to regulatory processes. Regulatory processes include: the modification of appraisal, e.g. by reappraising a situation; impulse control, e.g. the suppression of an action tendency; and the modification of action, e.g. by attenuating or replacing expressive behaviour.

Important for the instigation of regulatory processes are signals of aversive outcomes of unrestrained emotional behaviour. These outcomes can be external or internal. An example of an aversive external outcome is punishment, "when the environment retaliates, envies, disapproves, or despises because of emotions shown" (p. 409). Signals of aversive internal outcomes are "the calls of conscience and the sense of propriety" (p. 409).

In sum, emotional response is under dual control. Generative processes are modulated by regulatory processes. The next two sections highlight the importance of social norms for both emotion generation and emotion regulation.

3.4 The Role of Social Norms in Emotion Generation

Social norms enter the process of emotion generation during appraisal. The definition of the appraisal criterion "legitimacy" (see section 3.1) is based on social norms. Many emotions are contingent upon adherence or violation of social norms. Examples are "comfort in one's sense of propriety, pride in one's outstanding achievements, admiration for those of others; shame and guilt upon one's own infringements and distrust, anger, and indignation upon those of others" (p. 311). This list makes clear that to differentiate these emotions, the appraisal criterion "agency or responsibility" is necessary. Shame and guilt are contingent upon a norm violation by oneself, while contempt and anger are contingent upon a norm violation by another.

Scherer (1988, p. 112) provides a table of the complete appraisal patterns for some major emotions including shame, guilt, anger, contempt, and pride.

3.5 The Role of Social Norms in Emotion Regulation

Social norms are crucial for the instigation of emotion regulation. As mentioned in section 3.3, signals of aversive external or internal outcomes of unrestrained emotional behaviour instigate regulatory processes. Punishment was given as an example of an aversive external outcome. Of course, the violation of social norms is a major reason for punishment through sanctions.

Social norms also underly "the calls of conscience and the sense of propriety" signaling aversive internal outcomes. These signals consist in the anticipation of emotions such as shame or guilt that would be elicited by a norm violation.

Very important for the instigation of emotion regulation are social norms regarding the appropriateness of emotions and their expression. Hochschild (1983) focuses on culture-specific "feeling rules" and "expression rules." She shows that a good deal of our emotional life consists of "emotion work" that brings our emotions and their expression in line with these normative prescriptions. The rule against laughing at funerals mentioned in section 2 is an example of such a prescription.

Ekman and Friesen (1975) extensively discuss culturally defined "display rules" prescribing appropriate expressive behaviour. They distinguish four strategies for putting display rules into practice: "minimization," i.e., miniaturizing the expression; "maximization," i.e., exaggerating the expression; "masking," i.e., adopting a neutral expression; and "substitution," i.e., expressing a different emotion.

A considerable part of emotion socialization in childhood is devoted to the acquisition of norms regarding the appropriateness of emotions and their expression. Saarni (1993) distinguishes five methods of emotion socialization: direct instruction, contingency learning, imitation, identification with role models, and communication of expectancies.

4 An Emotion-Based View on an Influential Study

Conte and Castelfranchi (1995) realized that previous work in Artificial Intelligence (Shoham and Tennenholtz, 1992a,b) had a very restricted view of norms. Based on game theory, norms were seen essentially as conventions permitting or improving coordination among agents. Conte and Castelfranchi (1995) conducted a study to investigate another function of norms: the control of aggression among a population of agents. This research has been very influential, forming the basis of several studies (Walker and Wooldridge, 1995; Castelfranchi et al., 1998; Saam and Harrer, 1999). In the following, it is described briefly:

Agents perform some elementary routines for surviving in a situation of food scarcity (e.g., moving, eating, attacking an eating agent). Each agent has a strength, which is increased by eating and decreased by moving and attacking. In one condition, each agent owns a number of food items. All agents follow a normative strategy for aggression control: They do not attack agents eating their own food, i.e., they comply with the "finder-keeper" norm. In another condition, all agents follow a utilitarian strategy for aggression control: They do not attack eating agents whose strength is higher than their own. The normative strategy has been found to reduce aggression (i.e., the number of attacks) to a much greater extent than the utilitarian strategy.

Conte and Castelfranchi (1995) studied the function of the "finder-keeper" norm as a macro-social object. So the agents were deliberately kept as simple as possible and could just execute elementary routines. The term "aggression" simply denotes the execution of the "attack" routine.

How could agents be implemented that more accurately model the psychological processes underlying aggression control in humans? To this end, we point out that aggressive behaviour is a main example of emotional behaviour. Neither Conte and Castelfranchi (1995) nor the authors of the follow-up studies ever mention emotions.

Appraisal theory offers a detailed account of the processes underlying the generation and control of aggressive behaviour in humans:

Frijda calls the action tendency underlying aggressive behaviour "agonistic" (Frijda , 1986, p. 88). The agonistic action tendency covers attack and threat. The emotion corresponding to this action tendency is anger. The agonistic action tendency is generated by the appraisal that the satisfaction of a concern is obstructed. The end state of the agonistic action tendency is the removal of this obstruction.

A basic concern of a living being is the optimal state of feeding. Another person in possession of scarce food is appraised as obstructing the satisfaction of this concern. This appraisal leads to the generation of the agonistic action tendency. If this action tendency is not suppressed, overt aggressive behaviour (e.g. an attack) is generated.

Aggression control can thus be viewed as an example of impulse control, namely the suppression of the agonistic action tendency. In section 3.3 we mentioned that regulatory processes can be instigated by signals of aversive external or internal outcomes of unrestrained emotional behaviour. Punishment was given as an example of an external aversive outcome. When the person in possession of food is stronger than oneself, retaliation can lead to punishment for unrestrained aggression. If the "finder-keeper" norm is in force, aggression control is either due to the anticipation of punishment through sanctions or due to "the calls of conscience and the sense of propriety," i.e., the anticipation of shame or guilt as aversive internal outcomes (see section 3.5).

This short account of the generation and control of aggression suggests that appraisal theory can guide the development of a psychologically more plausible agent architecture. In the next section we will sketch our attempts to flesh out TABASCO[1], our appraisal-based agent architecture.

5 The Development of TABASCO

5.1 Existing Appraisal-Based Architectures

The majority of the current appraisal-based architectures used to engender emotional competence in software agents include some reified representation of a finite number of discrete emotional states through which all emotional processing is explicitly routed. Well-known examples of such architectures are the Affective Reasoner (Elliott, 1992) and Em (Reilly, 1996), the emotional component of the Tok architecture developed in the Oz project (Bates et al., 1992). Both architectures rely on the theory by Ortony et al. (1988), which quickly has become

[1] The name stands for "A Tractable Appraisal-Based Architecture for Situated Cognizers."

the most popular "reference model" of appraisal used in agent architectures.

The reification of emotional states results in the implementation of a full explicit mapping from these states to entailed effects, including internal processing and externally observable overt behaviour. The characteristics of systems engineered according to such a shallow approach are well known from the traditional research area of expert systems in artificial intelligence: the merits of rather straightforward design–for moderate ruleset size–and precisely known coverage stand against a number of problems, including brittleness of system behaviour surfacing with every occurrence of any situation not explicitly anticipated at implementation time; laboriousness of system extension; and the issue of overall system consistency: as the agent's behaviour is governed by a large collection of independent rules–as opposed to a small set of generating principles–it falls into the responsibility of designers and implementors to ensure that with a growing body of incorporated knowledge the system remains free of inconsistencies and continues to perform in a desired and coherent fashion.

Besides the problems of brittleness and consistency, reification of emotions as identifiable system components and routing of all processing through these entities engenders the problem of how to proceed from these emotions for further system processing, leading to the adoption of ad-hoc constructions of dubious validity.

The Affective Reasoner is an architecture for agents in a multi-agent world with the capability of abstract, domain-independent reasoning about emotion episodes. Such an architecture runs into serious problems when deployed in interactive virtual scenarios: to be effective in such applications, affective reasoning has to have appropriate access to pertinent information about and from the world, and has to be able to influence the overt external behaviour as well as the internal information processing of an agent. The only means to achieve this is to fully integrate emotional competence into an architecture which in turn has to be adapted to the environment in which the agent is situated.

5.2 The TABASCO Architecture

TABASCO is an attempt to overcome the problems stated above. It has first been adumbrated by Staller and Petta (1998). TABASCO integrates the three-level model of the appraisal process (see section 3.2) into a three-layer architecture for software agents situated in a virtual environment. Three-layer architectures have emerged as robust, widely adopted solutions to fundamental aspects of the realization of situated agents (Gat , 1997). Emotions are not modelled as reified entities, but as an adaptive process related to the agent-environment interaction, with the appraisal process and the execution of action tendencies as main components. Action tendencies provide a principled way of classifying the behavioural repertory of an agent in classes that share specific expressive characteristics, obviating the need of ad-hoc solutions. The implementation of the emotion process does not follow exactly Frijda (1986) who proposes a sequential process (see section 3.3). Our layered architecture has the advantage that the agent can respond reactively to events in the environment without having to execute a sequential process with action generation as the last step.

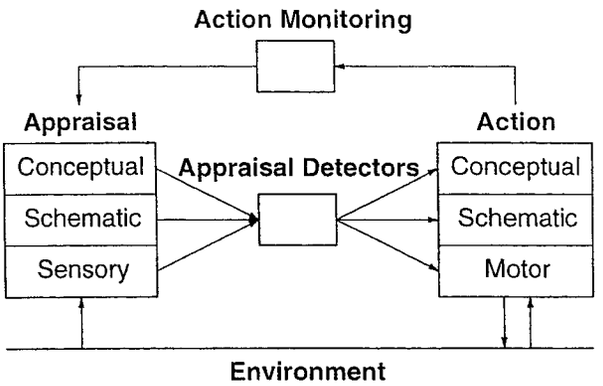

Figure 1: The TABASCO Architecture

In the following, we sketch the conceptional design, which is shown in figure 1.

The psychological idea underlying TABASCO is that the distinction between sensory-motor, schematic, and conceptual processing does not only apply to appraisal, but also to the generation of action, as proposed originally by Leventhal in his "perceptual-motor theory of emotion" (Leventhal, 1984).

The **Appraisal** component processes environmental stimuli and models the appraisal process based on the three-level theory of the appraisal process (see section 3.2).

The **Action** component models long-term planning processes at the conceptual level, the generation of action tendencies at the schematic level, and action generation at the motor level.

The **Appraisal Detectors**, suggested by Smith and Kirby (2000), detect and combine the appraisal outcomes, and instigate processes in the **Action** component: planning processes, action tendencies, and actions.

The **Action Monitoring** component monitors the planning and execution processes in the **Action** component and sends the results to the **Appraisal** component, where it is integrated into the appraisal process.

So far we have mainly been concerned with designing an architecture for emotion generation. The whole range of regulatory processes described by Frijda (1986) has not yet been incorporated into the TABASCO architecture. But e.g. impulse control can be modelled by letting the planning processes at the conceptual level intervene in the processes at the schematic level so that action tendencies can be prevented from being executed. It is also possible that results of **Action Monitoring** that are

sent to the **Appraisal** component lead to a reappraisal of a situation. For example, the execution of actions without success may lead to a reappraisal of the appraisal criterion "perceived control, power or coping potential" (see section 3.1).

A version of TABASCO has been implemented for the control of a synthetic character interacting with users in an immersive interactive virtual environment (Petta, 1999; Petta et al., 1999). The implementation is based on 3T (Bonasso et al., 1999), a three-layer agent architecture with the layers *deliberation*, *sequencing*, and *reactive skills*. The deliberation layer consists of a planner and corresponds to conceptual processing. The sequencing layer corresponds to schematic processing. It is based on the Reactive Action Packages (RAPs) system. (Firby, 1989). The reactive skills correspond to sensory-motor processing. So far we have concentrated on implementing the generation and management of action tendencies based on RAPs.

Another line of research along which we are trying to flesh out TABASCO is the implementation of FORREST (Petta et al. , 2000), an agent situated in multi-user real-time text-based environments known as MUDs (Curtis , 1992). FORREST is an expansion of the Colin MUD-bot (Mauldin , 1994). The C code of Colin was complemented with a fairly accurate implementation of Frijda's sequential model of the emotion process (see section 3.3). Most of the emotion process takes place in a module written in NASA's expert system programming shell, CLIPS (1993). The rule-based implementation of a sequential emotion process forms the basis of the conceptual level of TABASCO. The next step towards a realization of TABASCO is the implementation of associative processing at the schematic level. With respect to social simulations, a MUD has the advantages that it is already designed as a multi-user system, in which an arbitrary number of agents and users can interact with each other and share equal "symbolic" access to the environment. We plan to exploit these facts in future social simulations.

Social norms have not yet been incorporated in our implementation. In the next section, we present some first theoretical considerations for the incorporation of social norms into TABASCO.

5.3 Incorporating Social Norms into TABASCO

5.3.1 Emotion Generation

In section 3.4 we pointed out that social norms enter the process of emotion generation during appraisal. For the evaluation of the "legitimacy" criterion it must be determined whether a social norm has been violated.

The implementation of the normative reasoning processes underlying the "legitimacy" check at the conceptual level of the **Appraisal** component can certainly profit by research on deontic logic. For example, Conte et al. (1999) present a logical framework for the specification of "norm-autonomous" agents. Their approach is based on explicit representations for goals, intentions, and beliefs. A norm is conceptualized as an obligation on a given set of agents to accomplish or abstain from a given action. Incorporating norms into agent architectures based on logic is a common approach. However, it would be problematic to base an architecture for a situated agent solely on a logical framework. Wooldridge and Jennings (1995) point out several problems of such a "deliberative" agent architecture, e.g., the problem of maintaining an explicitly represented, symbolic world model in a rapidly changing environment. In contrast, a layered architecture such as TABASCO allows the agent to react directly to changes in the environment without relying on a world model.

In TABASCO the evaluation whether a norm has been violated is not restricted to the conceptual level. The schematic level is also involved. It has even been hypothesized by Leventhal and Scherer (1987) and proponents of other multi-level theories (Teasdale , 1999) that the associative processes at the schematic level are necessary for emotion elicitation, while the "cold" cognitions at the conceptual level have only a subsidiary function. In the following, we present some theoretical ideas on the incorporation of social norms into the schematic level of the **Appraisal** component:

Leventhal and Scherer (1987) use social schemas for the conceptualization of social norms at the schematic level, but do not provide any detail. Social schemas are a central concept of social cognition (Fiske and Taylor , 1991; Augoustinos and Walker , 1995).

Important social schemas are event schemas (scripts), which specify the appropriate behavioural sequence of events, e.g. of eating in a restaurant. Scripts were introduced by Schank and Abelson (1977) to account for the human ability to understand more than was being referred to explicitly in a sentence by explaining the organization of implicit knowledge of the world one inhabits. When we hear the sentence "John ordered sushi but he didnt like it," the restaurant script allows us to infer that this sentence is about eating.

Schank (1982) modified his view of scripts. The starting point of his theory is the conceptualization of a script as a dynamic memory structure. A script is not an unchangeable data structure, but changes over time by storing the memories of episodes deviating from the script. For example, a person who has never been in a Japanese restaurant uses the restaurant script to form expectations about what will happen. Receiving chopsticks instead of a fork is an expectation failure. This expectation failure is stored at the script juncture where it occurred, so that the next time the person receives chopsticks the memory is retrieved and made available for use. Schank calls this conception of memory failure-driven memory.

This conception of scripts is useful for implementing social norms and the detection of norm violations at the schematic level of the **Appraisal** component. The script defines the sequence of actions prescribed by the norm, while the episodes deviating from the script correspond to norm violating episodes. So the detection of a norm violation simply amounts to reminding of expectation failures.

In fact, the implementation of the schematic level of the **Appraisal** component can generally be based on the conception of scripts as organizing memories of expectation failures. Unexpected events are exactly the kind of events that can give rise to emotions. Based on the model by Smith and Kirby (2000) (see section 3.2), the memories of these events are directly associated with the respective appraisal patterns. Then appraisal at the schematic level merely involves reminding deviations from the script and following the link to the associated appraisal pattern.

Schank (1982) further elaborates his theory based on so-called memory organization packets (MOPs). MOPs cover more general knowledge than scripts. For example, Schank proposes the existence of a MOP for a professional office visit that applies to visits to a doctor and to a lawyer equally, while these events are covered by separate scripts. This theory has formed the basis of case-based reasoning (Kolodner , 1993) and can account for more results of memory research than scripts alone. The final implementation of the schematic level of the **Appraisal** component may be based on this theory and case-based reasoning techniques, e.g. for case representation and indexing.

5.3.2 Emotion Regulation

In section 3.5 we pointed out that social norms play an important role in emotion regulation because a norm violation through unrestrained emotional behaviour can be the reason for punishment (an aversive external outcome) or for the generation of emotions such as shame and guilt (an aversive internal outcome). Crucial for the instigation of regulatory processes is the ability to anticipate these aversive outcomes. This ability largely relies on learning. For example, if a certain emotional behaviour has led to negative consequences, a memory of this experience can be stored in memory and used for the timely instigation of regulatory processes in similar situations in the future.

How can such a memory-based instigation of regulatory processes be modelled in TABASCO? As an example, we focus on impulse control based on the memory that a previous execution of an action tendency led to guilt. We cannot specify the computational processes exactly, but we outline which components of TABASCO may be involved in an implementation:

In section 5.2 we suggested that impulse control can be modelled by letting processes at the conceptual level of the **Action** component prevent action tendencies generated at the schematic level from being executed. In order to suppress an action tendency, the conceptual level of the **Action** component must have access to the memory of a similar situation in which the action tendency was executed. This memory is located at the schematic level of the **Appraisal** component and may be represented and retrieved based on Schank's (1982) memory theory or case-based techniques. The association of this situation with guilt is represented by means of an associative link between the memory of the situation and the appraisal pattern of guilt. Currently, the design as shown in figure 1 does not contain a direct connection between the schematic level of the **Appraisal** component and the conceptual level of the **Action** component, but there is no reason against it.

Our emotion-based view of the study by Conte and Castelfranchi (1995) presented in section 4 suggests that appraisal theory could guide the development of agents that model the processes of aggression control in a psychologically more plausible way. In TABASCO the behaviour of agents complying with the "finder -keeper" norm could be modelled by the processes of memory-based impulse control outlined above. The action tendency to be suppressed is the agonistic action tendency.

Our conception of a memory-based instigation of regulatory processes in TABASCO is a way of modelling what Frijda refers to as "the calls of conscience and the sense of propriety" (see section 3.3). Even if no external punishment is expected, regulatory processes are instigated based on memories of situations associated with appraisal patterns of guilt or shame.

Emotion regulation based on memories associated with outcomes of unrestrained emotional behaviour is an instance of contingency learning, which has been identified by Saarni (1993) as a mechanism of emotion socialization (see section 3.5). The other mechanisms such as direct instruction, imitation, identification with role models, and communication of expectancies certainly require more complex cognitive processes, and proposing how to model them is beyond the scope of this paper.

6 Functions of Emotions at the Macro Level

Emotions have an important adaptive function for the individual. According to appraisal theory, they support the individual in the satisfaction of concerns or goals by instigating action tendencies. These action tendencies are directed towards establishing or maintaining a certain relationship with the environment. However, the environment is a social environment. Appraisal theory focuses on the internal psychological processes underlying emotions, but remains largely silent about potential social functions of emotions.

In this section we briefly review three theories suggesting that emotions have the important function of sus-

taining norms of cooperation and reciprocity. These theories do not explicitly claim that emotions sustain social norms, but they share the view that certain emotions (e.g., anger) bring a person to punish a cheater (i.e., a person who failed to cooperate or reciprocate). Under the assumption that norms of cooperation and reciprocity are in force, this amounts to imposing a sanction on a norm violator. Regarding the existence of such norms, it has been hypothesized that the norm of reciprocity is universal, i.e., that it exists in all human cultures (Gouldner , 1960).

6.1 Reciprocal Altruism

Altruistic behaviour benefits another person, while being apparently detrimental to the person performing the behaviour. Helping and sharing food are examples of altruistic behaviour. Trivers (1971) explains altruistic behaviour toward nonkin with a theory of "reciprocal altruism." Based on this theory, people perform an altruistic act in the expectation that the beneficiary will reciprocate in the future.

Trivers argues that emotions play a crucial role in the evolution of reciprocal altruism. For example, "moralistic aggression" has been selected for in order to punish unreciprocating individuals ("cheaters") e.g. by cutting off all future altruistic acts. Guilt has been selected for in order to motivate the cheater to make up for his misdeed and thus to continue reciprocal relationships. Trivers enumerates a number of other emotions that he regards as important for the regulation of the altruistic system.

6.2 Emotions as "Commitment Operators"

Based on Trivers's work, Aubé (1998) proposes that emotions might have emerged to control and manage commitments among members of a society. Aubé borrows the notion of commitment from symbolic interactionism in sociology (e.g., Becker, 1960) and distributed artificial intelligence (e.g., Fikes, 1982). Commitments bind agents together into cooperative behaviour. Aubé calls emotions "commitment operators" that "operate so as to establish or create new commitments (joy, gratitude), protect, sustain, or reinvest old ones (joy, hope, gratitude, pride), prevent the breaking of commitments by self or others (pride, guilt, gratitude, anger), or call on 'committed' others in cases of necessity, danger, and helplessness (sadness, fear)." (Aubé, 1998, p. 15).

Commitments are conceived of as "second-order resources" in addition to vital "first-order resources" such as food. Based on this classification of resources, Aubé suggests a two-layer control system: Needs such as hunger control first-order resources, while emotions control second-order resources.

6.3 Emotions as "Commitment Devices"

Frank (1988) also uses the term "commitment," but his conception of this term differs from Aubé's. Frank proposes that in social dilemmas such as the prisoner's dilemma some emotions, the so-called "moral sentiments," commit a person to act contrary to self-interest. For example, the predisposition to feel guilt commits a person to cooperate, even if cheating were in his material interest. A person with the predisposition to get outraged after having been cheated is committed to punish the cheater, even if it is costly in material terms. So emotions such as guilt and anger act as "commitment devices" that change the material incentives.

But there must be a material gain from having these emotions, otherwise they would not have evolved. Frank proposes that emotional predispositions have long-term material advantages: An honest person with the predisposition to feel guilt will be sought as a partner in future interactions. The predisposition to get outraged will deter others from cheating.

However, others must be able to discern the presence of these emotional predispositions. Frank suggests two ways how this might occur: The first is reputation. The knowledge about the honesty or the vengefulness of a person can be spread among the population. The second way of discerning emotional predispositions is through physical and behavioural clues, such as facial expressions, voice, and posture. Frank discusses the reliability of these clues and the problem of deception, but this discussion is beyond the scope of this paper.

7 Connecting the Macro and Micro Levels

The theories reviewed above focus on the functions of emotions at the macro level, while appraisal theory specifies the processes occurring at the micro level. Is there any connection between these two levels of analysis? Indeed, the macro-level theories make assumptions about micro-level processes that are fully in accordance with appraisal theory.

For example, the theories assume that the experience of having been cheated leads to anger. But what is the psychological process underlying the realization that one has been cheated? It can be thought of as an appraisal process: Having been cheated is appraised as a situation in which the satisfaction of a concern or goal has been obstructed and another agent is responsible for this obstruction. These are the crucial appraisal outcomes for the generation of anger.

The theories also assume that emotions have an influence on actions. For example, Trivers claims that guilt motivates the cheater to make up for his misdeed. This is exactly the action tendency of guilt. Punishing a cheater can be interpreted as due to the agonistic action tendency

of anger.

These examples suggest that appraisal theory can serve as a link between the macro and micro levels. Macro-level functions of emotions such as sustaining cooperation and reciprocity depend on the micro-level processes of appraisal and the generation of action tendencies. This insight paves the way for testing the macro-level theories in social simulations with agents that are able to perform appraisal and the generation of action tendencies. TABASCO is a proposal for the implementation of such agents.

8 Conclusion

In this paper, we have tried to show that the computational study of social norms can profit by modelling emotions among agents in artificial societies. We have suggested appraisal theory as the theoretical foundation for endowing agents with emotions. Our TABASCO architecture is a proposal for the development of appraisal-based agents. The computational study of social norms can benefit from the development of TABASCO in the following ways:

- Social norms must be represented in TABASCO. Appraisal theory, especially the three-level theory of the appraisal process can guide the exploration of representations that are not based on logic. We have suggested social schemas, especially scripts, as the basis for this exploration.
- The insight that appraisal and action tendencies can serve as a link between the macro and micro levels paves the way for testing the macro-level emotion theories – which suggest that emotions serve to sustain norms of cooperation and reciprocity – in social simulations with TABASCO agents.
- The account of appraisal theory for emotion regulation through social norms sheds new light on existing research. From the point of view that aggression control is an instance of impulse control, a large part of computational research on social norms has investigated a special case of emotion regulation through social norms. The implementation of regulatory processes in TABASCO leads to a psychologically plausible model of emotion regulation through social norms.

In general, emotions are of paramount importance for the social life of humans and should therefore not be neglected in the study of artificial societies.

Acknowledgements

We would like to thank Robert Trappl for helpful comments on an earlier draft of this paper. This research has been supported by the Austrian Federal Ministry of Science and Transport.

References

M. Aubé. A commitment theory of emotions. In *Emotional and Intelligent: The Tangled Knot of Cognition, Proceedings of the 1998 AAAI Fall Symposium*, AAAI Technical Report FS-98-03, Orlando, FL, 13–18, 1998.

M. Augoustinos and I. Walker. *Social Cognition: An Integrated Introduction*. Sage, London, 1995.

J. Bates, A.B. Loyall, and W.S. Reilly. An architecture for action, emotion, and social behavior. In *Proceedings of the Fourth European Workshop on Modeling Autonomous Agents in a Multi-Agent World*. San Martino al Cimino, Italy, 1992.

H.S. Becker. Notes on the concept of commitment. *American Journal of Sociology*, 66:32–40, 1960.

R.P. Bonasso, R.J. Firby, E. Gat, D. Kortenkamp, D.P. Miller, and M. Slack. Experiences with an architecture for intelligent, reactive agents. In H. Hexmoor (ed.), Special Issue: Software Architectures for Hardware Agents, *Journal of Theoretical and Experimental Artificial Intelligence*, 9(2/3):237–256, 1997.

C. Castelfranchi, R. Conte, and M. Paolucci. Normative reputation and the costs of compliance. *Journal of Artificial Societies and Social Simulation*, 1(3), <http://www.soc.surrey.ac.uk/JASSS/1/3/3.html>, 1998.

R. Conte and C. Castelfranchi. Understanding the functions of norms in social groups through simulation. In G.N. Gilbert and R. Conte (eds.), *Artificial Societies: The Computer Simulation of Social Life*. UCL Press, London, 252–267, 1995.

R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In J.P. Müller, M.P. Singh, A.S. Rao (eds.), *Intelligent Agents V - Proceedings of the Fifth International Workshop on Agent Theories, Architectures, and Languages (ATAL- 98)*. Springer-Verlag, Heidelberg, 99–112, 1999.

P. Curtis. Mudding: Social phenomena in text-based virtual realities. In *Proceedings of the 1992 Conference on Directions and Implications of Advanced Computing*, Berkeley, CA, 1992.

P. Ekman and W.V. Friesen. *Unmasking the Face*. Prentice Hall, Englewood Cliffs, NJ, 1975.

C.D. Elliott. *The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System*. PhD Thesis, Northwestern University, Evanston, IL, 1992.

J. Elster. *The Cement of Society: A Study of Social Order*. Cambridge University Press, Cambridge, UK, 1989.

J. Elster. Rationality and the emotions. *The Economic Journal*, 106(438):1386–1397, 1996.

J. Elster. *Alchemies of the Mind: Rationality and the Emotions*. Cambridge University Press, Cambridge, UK, 1999.

R.E. Fikes. A commitment-based framework for describing informal cooperative work. *Cognitive Science*, 6(4):331–348, 1982.

R.J. Firby. *Adaptive Execution in Complex Dynamic Worlds*. PhD Thesis, Yale University, New Haven, CT, 1989.

S.T. Fiske and S.E. Taylor. *Social Cognition* (2nd edition). McGraw-Hill, New York, 1991.

R.H. Frank. *Passions within Reason: The Strategic Role of the Emotions*. Norton, New York, 1988.

N.H. Frijda. *The Emotions*. Cambridge University Press, Cambridge, UK, 1986.

N.H. Frijda. The place of appraisal in emotion. *Cognition and Emotion*, 7(3/4):357–388, 1993.

N.H. Frijda, P. Kuipers, and E. ter Schure. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, 57(2):212-228, 1989.

E. Gat. On three-layer architectures. In D. Kortenkamp, R.P. Bonasso, and R. Murphy (eds.), *Artificial Intelligence and Mobile Robots*, MIT/AAAI Press, 1997.

J.J. Gibson. *The Ecological Approach to Visual Perception*. Houghton-Mifflin, Boston, 1979.

A.W. Gouldner. The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25:161–178, 1960.

A.R. Hochschild. *The Managed Heart: Commercialization of Human Feeling*. University of California Press, Berkeley, 1983.

J. Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, San Mateo, 1993.

R.S. Lazarus. *Emotion and Adaptation*. Oxford University Press, New York, 1991.

J.E. LeDoux. *The Emotional Brain*. Simon & Schuster, New York, 1996.

H. Leventhal. A perceptual-motor theory of emotion. *Advances in Experimental Social Psychology*, 17:117–182, 1984.

H. Leventhal and K.R. Scherer. The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition and Emotion*, 1(1):3–28, 1987.

M.L. Mauldin. CHATTERBOTS, TINYMUDS, and the Turing Test: Entering the Loebner Prize Competition. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*. AAAI Press/MIT Press, Cambridge, MA, 16–21, 1994.

L.Z. McArthur and R.M. Baron. Toward an ecological theory of social perception. *Psychological Review*, 90(3):215–238, 1983.

A. Ortony, G.L. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge, UK, 1988.

P. Petta. Principled generation of expressive behavior in an interactive exhibit. In J.D. Velasquez (ed.), *Workshop: "Emotion-Based Agent Architectures" (EBAA'99)*, Third International Conference on Autonomous Agents (Agents '99), Seattle, WA, 94–98, 1999.

P. Petta, A. Staller, R. Trappl, S. Mantler, Z. Szalavari, T. Psik, and M. Gervautz. Towards engaging full-body interaction. In H.-J. Bullinger and P.H. Vossen (eds.), *Adjunct Conference Proceedings of the 8th International Conference on Human-Computer Interaction (HCI International '99) jointly with the 15th Symposium on Human Interface (Japan)*. Fraunhofer IRB Verlag, Stuttgart, 280–281, 1999.

P. Petta, M. Macmahon, A. Staller. FORREST: Forschung ueber/research on emotion simulation. In C. Landauer and K.L. Bellman (eds.), *Proceedings of the Virtual Worlds and Simulation Conference*, 2000 Western Multiconference. Society for Computer Simulation International, San Diego, CA, 2000.

C.M. van Reekum and K.R. Scherer. Levels of processing in emotion-antecedent appraisal. In G. Matthews (ed.), *Cognitive Science Perspectives on Personality and Emotion*. Elsevier, Amsterdam, 259–300, 1997.

I.J. Roseman. Cognitive determinants of emotion: A structural theory. In P. Shaver (ed.), *Review of Personality and Social Psychology* (Vol. 5). Sage, Beverly Hills, CA, 11–36, 1984.

N.J. Saam and A. Harrer. Simulating norms, social inequality, and functional change in artificial societies. *Journal of Artificial Societies and Social Simulation*, 2(1), <http://www.soc.surrey.ac.uk/JASSS/2/1/2.html>, 1999.

C. Saarni. Socialization of emotion. In M. Lewis and J.M. Haviland (eds.), *Handbook of Emotions*. Guilford Press, New York/London, 435-446, 1993.

R.C. Schank. *Dynamic Memory - A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, Cambridge, UK, 1982.

R.C. Schank and R. Abelson. *Scripts, Plans, Goals and Understanding*. Erlbaum, Hillsdale, 1977.

K.R. Scherer. On the nature and function of emotion: A component process approach. In K.R. Scherer and P. Ekman (eds.), *Approaches to Emotion*. Erlbaum, Hillsdale, NJ, 293–318, 1984.

K.R. Scherer. Criteria for emotion-antecedent appraisal: A review. In V. Hamilton, G.H. Bower, and N.H. Frijda (eds.), *Cognitive Perspectives on Emotion and Motivation*. Kluwer, Dordrecht, 89–126, 1988.

K.R. Scherer. Appraisal theory. In T. Dalgleish and M. Power (eds.), *Handbook of Cognition and Emotion*. Wiley, Chichester, 637–663, 1999.

Y. Shoham and M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies (preliminary report). In *Proceedings of the Tenth National Conference on Artificial Intelligence*. AAAI Press/MIT Press, Cambridge/Menlo Park, 276–281, 1992.

Y. Shoham and M. Tennenholtz. Emergent conventions in multi-agent systems: Initial experimental results and observations (preliminary report). In *Proceedings of the Third International Conference on Knowledge Representation and Reasoning*. Kaufman, San Mateo, 225–231, 1992.

C.A. Smith and P.C. Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48:813–838, 1985.

C.A. Smith and L.D. Kirby. Affect and appraisal. In J.P. Forgas (ed.), *Feeling and Thinking: The Role of Affect in Social Cognition*. Cambridge University Press, Cambridge, UK, 2000.

C.A. Smith and R.S. Lazarus. Emotion and adaptation. In L.A. Pervin (ed.), *Handbook of Personality: Theory and Research*, Guilford, New York, 609–637, 1990.

A. Staller and P. Petta. Towards a tractable appraisal-based architecture for situated cognizers. In D. Cañamero, C. Numaoka, and P. Petta (eds.), *Grounding Emotions in Adaptive Systems*, Workshop Notes, 5th International Conference of the Society for Adaptive Behaviour (SAB'98), Zurich, Switzerland, 56–61, 1998.

J.D. Teasdale. Multi-level theories of cognition-emotion relations. In T. Dalgleish and M. Power (eds.), *Handbook of Cognition and Emotion*. Wiley, Chichester, 665–681, 1999.

R.L. Trivers. The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46:35–57, 1971.

A. Walker and M.J. Wooldridge. Understanding the emergence of conventions in multi-agent systems. In *Proceedings of the First International Conference on Multiagent Systems (ICMAS'95)*. AAAI Press, San Francisco, 1995.

M.J. Wooldridge and N.R Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

R.B. Zajonc. Feeling and thinking: Preferences need no inferences. *American Psychologist*, 2:151–176, 1980.

CLIPS Reference Manual: Volumes I&II, Lyndon B. Johnson Space Center, Software Technology Branch, 1993.

The Society of Mind Requires an Economy of Mind[1]

Ian Wright
Sony Computer Entertainment Europe
Waverley House, 7-12 Noel Street, London, W1V 4HH
ian_wright@scee.net

Abstract

A society of mind will require an economy of mind, that is multi-agent systems that meet a requirement for the adaptive allocation and reallocation of scarce resources will need to employ a quantitative, universal, and domain-independent representation of value that mirrors the flow of agent products, much as money is used in simple commodity economies. The money-commodity in human economic systems is shown to be an emergent exchange convention that serves both to constrain and allow the formation of commitments by functioning as an ability to buy processing power. Multi-agent systems with both currency flow and minimally economic agents can adaptively allocate and reallocate control relations and scarce resources, in particular labour or processing power. The implications of this design hypothesis for cognitive science and economics are outlined.

1 The society of mind

> "... a group of agencies inside the brain could exploit some 'amount' to keep account of their transactions with one another. Indeed agencies need such techniques even more than people do, because they are less able to appreciate each other's concerns. But if agents had to 'pay their w..y', what might they use for currency? One family of agents might evolve ways to exploit their access to some chemical that is available in limited quantities; another family of agents might contrive to use a quantity that doesn't actually exist at all, but whose amount is simply 'computed'."

M. Minsky, The Society of Mind, "magnitude and marketplace", page 284.

Marvin Minksy's The Society of Mind (Minsky, 1987) is the best example of the social metaphor applied to the understanding and design of minds. It outlines a computational society of heterogeneous agents that compete and co-operate to produce mental capabilities. The approach of decomposing a computational mind into a society of less intelligent agents is compelling because social systems and large, parallel computing systems share design features. For example, both kinds of system consist of a set of mutually connected, interacting subcomponents that are able to perform work, such as computational units that process or people that labour. Computational and human agents function as both producers and consumers, for example the input and output of information or the consumption and production of commodities. Agents may perform specialist functions, such as modular decomposition in software systems and the division of labour in social systems. Agents may operate in parallel, that is subcomponents may function relatively autonomously and concurrently perhaps pursuing their own local goals. Both computational and human societies need to be co-ordinated by mechanisms for the production, distribution and consumption of agent products, such as globally accessible databases or free market mechanisms. In addition, these systems must adaptively allocate scarce resources, be they limited labour resources or processing time or commodities or information in restricted supply.

These considerations suggest that the social "metaphor" is no metaphor at all, but is a partial identity between a class of complex systems at the information processing level of abstraction. However, as with all compelling parallels it is important to identify differences as well as similarities. Furthermore, it would be a mistake to base computational theories on current ideas about social organisation given the current dominance of social empiricism and the lack of a science of social design. Despite these warnings, the aim of this paper is to argue that a society of mind will require an economy of mind, and that economic theories,

[1] This is a revised version of (Wright & Aube, 97).

concepts and methods have new applications in multi-agent systems (MAS) and the understanding of cognition. The paper, therefore, emphasises similarities not differences, and is primarily speculative, bearing on the foundations of adaptive multi-agent systems.

The key idea is that a quantitative, universal, and domain-independent representation of value, exchanged between mental subcomponents much like money is exchanged in human economies, is necessary for the satisfaction of certain design requirements for both natural and artificial adaptive minds. Minsky anticipated such an idea, and this paper attempts to develop it further.

2 The co-ordination problem in multi-agent systems

A MAS can be thought of as a system that is composed of a collection of agents that normally have their own beliefs and goals, sharing a domain that allows actions to be performed, including communicative actions, such that the system meets some global requirements. The global requirements normally specify goals that can be met by agents acting co-operatively, competitively or both to discover solutions. Jennings (Jennings, 1996) discusses the co-ordination problem in MAS, which is the problem of ensuring that a society of agents interact in such a manner to achieve global goals given available resources. Co-ordination is required because "there are dependencies between agent actions", "there is a need to meet global constraints" and "no one individual has sufficient competence, resources or information to solve the entire problem". Without co-ordination the MAS would fail to produce useful global results. Jennings states that all co-ordination mechanisms can ultimately be reduced to commitments and their associated conventions. Commitments need not be generated in a conscious and deliberative manner: attachment structures in most bird and mammal species, for instance, involve some kind of built-in commitments already "installed" between certain individuals (selective mating, caring and protection of the young, territorial defense and so forth), without necessarily relying upon a conscious and explicit contractual basis.

A commitment is essentially a goal: an agent can make a commitment to itself (e.g., "I will tidy my desk today") or to others, in which case it can be thought of as a pledge or promise (e.g., "I will meet you at ten tomorrow"). As goals, commitments could result from many goal generators, some very primitive, and some more deliberative. Joint commitments are possible (e.g., "We will both move house") and are preconditions for co-operative action. Conventions manage commitments: they are rules that determine how an agent's commitments are to be formed, reconsidered, or rejected; and social conventions are rules that determine how agents should behave towards each other, for example if they change mutual commitments. For example, agent A may commit to meet agent B at ten because it is conventional for A to obey B because B has greater authority. Subsequently, however, A acquires a more pressing commitment and does not have sufficient time resources to honour the commitment to B. Hence, A informs B of the difficulty because it is a social convention to do so, allowing B to re-plan and ask another agent C, who can do the work of A, to meet at ten. This is an example of co-operation, communication of failure and re-planning. Designing conventions and social conventions is difficult. It is likely that in natural systems, powerful mechanisms have evolved to generate, protect, manage and regulate conventions (Aube & Senteni, 1996a; Aube & Senteni,1996b). Designing conventions amounts to designing a set of rules that can interact to produce coherent and useful emergent behaviour.

In summary, MAS need to be co-ordinated if they are to meet overall goals. Commitments and conventions can achieve co-ordination.

3 Adaptive multi-agent systems

Adaptive multi-agent systems (AMAS) are a type of MAS that can continually reconfigure their activity to produce solutions that meet changing global requirements. The class of AMAS is sufficiently general to include many diverse kinds of system and mechanism, in much the same way as the class of adaptive agent architectures can include such mechanisms as reinforcement learning algorithms, artificial neural networks, genetic algorithms, and so forth. In the abstract, there are three distinct ways in which an AMAS can modify its global behaviour. It can (i) alter the behaviours of individual agents or (ii) alter the control relations between agents, for example dynamically defining groups of leader and follower agents. (i) is a change of commitments, and (ii) is a change in conventions and social conventions. Alternatively, (iii) existing agents may be removed or qualitatively new and behaviourally different agents may be introduced into the system. AMAS require co-ordination mechanisms that can cope with this kind of changing complexity. Such mechanisms need to allocate and reallocate agents to different tasks, alter social hierarchies, change individual agent behaviours to fit new circumstances, and provide means by which global

constraints can direct local processing without the need for high bandwidth communication. In addition, there need to be natural ways in which global constraints can be defined within the system.

(i), (ii) and arguably (iii) all occur in a natural, adaptive multi-agent system. How then is co-ordination achieved in human society?

4 Money and exchange-value

Human designers of robots often turn to the natural world for design ideas. Similarly, human designers of co-ordination mechanisms for MAS can also turn to the natural world. The study of ant colonies, primate groups and human social interaction are all potential sources of inspiration. For example, (Aube & Senteni, 1996a; Aube & Senteni, 1996b) propose that the emotions arose to co-ordinate animal groups and therefore can serve as a foundation for co-ordination in MAS. They view commitments as a special kind of resource that ensures access to basic commodities of survival value, and emotional structures as the control mechanism that manages these special resources. This section develops the contention that human economic activity provides an example of another important co-ordination mechanism - currency flow - that may be common to a certain class of adaptive MAS. We even think that such a view might help uncover the inner mechanics of motivations: that is, why and how some mental processes within the society of mind come to take precedence (be "preferred") over others.

4.1 Fundamental requirements for the development of money

All human societies are in commerce with nature, extracting raw materials from the environment and returning human waste to the earth. Social organisation implies a division of labour amongst the individuals of the society, that is individuals perform different, socially useful functions. The total labour of society is shared between the different functions, and the products of this labour distributed according to some, usually implicit, scheme and through some collection of mechanisms. One very obvious requirement for a successful social system is that it reproduce its conditions of existence; that is, it must create conditions such that individuals survive and produce offspring. This requirement entails that what is produced, distributed and consumed should be so organised to satisfy those needs. This is one of the important co-ordination problems that social organisations are required to solve: labour must be divided and its products distributed so that at least a sufficient number of individuals' basic needs are met. This defines a major global constraint for successful human social systems.

Money arose at a certain point in human history to solve problems of production, consumption and exchange. Pure gold was first coined as money in 625 BC in Greece (Boardman, Griffin & Murray, 1993). In a matter of fifty years trade had burgeoned, and banks, merchants, and moneylenders appeared. A numerical representation of value had a revolutionising effect on the capabilities of human society. Subsequently, currency flow has been a common feature of human social organisation, surviving and developing through classical society, feudal arrangements, and industrial and modern finance capitalism. To understand the function of money it is necessary to examine how and why it arose. The following account of the development of money is based on the opening analysis in Marx's Capital (Marx, 1970). It is a rough historical sketch of the emergence of a social convention in human society. The account abstracts from the real historical development of money and uses simple stages and examples for the purpose of exposition. In addition, the emergence of money is examined in an idealised simple commodity economy, allowing later complications such as price-fixing, cartels, monopolies, taxation, trade tariffs, transportation costs, power relations, trade unions, and the legislative power of the state, to be ignored.

Stage one - simple exchange or swapping. Individual and relatively self-sufficient producers with a small surplus product, such as a peasant farmer, whose chickens have lain too many eggs, exchange their goods for other goods. For example, 24 eggs may be exchanged for 2 loaves of bread. In this isolated act of exchange the equality relation (24 eggs = 2 loaves) is determined by the producers' respective opinions of the use-value of the other's goods. The term "use-value" simply means that the good satisfies some desire or need. In other words, the respective values of the goods are determined locally and subjectively. The exchange of products has a precondition: each producer must have a surplus-product that the other desires. All exchange is performed with a view to obtaining another's surplus-product for the purposes of consumption. Money does not as yet exist.

Stage two - extended exchange or organised swapping. The development of better production techniques and increase in population size creates a greater surplus product available for exchange. Instead of isolated acts of exchange there may be a definite geographical locale

where trading takes place, that is the market. The peasant's 24 eggs now enter into potential relations with all the other commodities available. For example, the 24 eggs may now be exchanged for 2 loaves, or a pair of boots, or five candles, or a pound of butter and so forth. Importantly, an element of competition appears that was not present in stage one. Instead of a single peasant and consumer there is a social community of interconnected producers and consumers, for example peasants, bakers, and candlestick makers. Given the choice a baker will tend to exchange his bread for as many eggs as he can get from the community of peasants; conversely, a peasant will tend to exchange his eggs for as many loaves as he can get from the community of bakers. This systemic dynamic - colloquially, the notion of "shopping around" - will, all other things being equal, have a tendency to force the equivalence relation between eggs and bread towards a particular ratio that holds for all such transactions. This equivalence relation will thus be determined by the joint action of the peasants and bakers. The respective values of the commodities are now determined globally and socially as opposed to locally and subjectively in stage one. An individual's local calculation increasingly becomes ineffective in the determination of the equivalence relation, which now tends to be fixed by the community as a whole.

Stage three - ubiquitous exchange. A community in which a good deal of exchange occurs soon finds it convenient to select a particular commodity to serve as the general form of value. A widely valued article would be the commodity to choose. This special commodity then serves as a unit of comparison of value and is directly exchangeable with all other commodities, thereby overcoming the limitations of organised swapping, as all producers are now willing to swap their goods for the general form of value. There need be no local coincidence of wants.

Stage four - money. As soon as a particular commodity is socially agreed upon to serve as the general form of value it becomes the money-commodity, that is it serves as a universal means of exchange. In most societies this commodity has been gold or silver, and not cows. For example, if 24 eggs = 1 measure of gold, and 1 measure of gold is coined as 10 pence, then 24 eggs have the price 10p. Gold can serve as the embodiment of value, and may be exchanged for any other commodity. "Although gold and silver are not by Nature money, money is by Nature gold and silver" (Marx, 1970). Precious metals were chosen because they exhibit uniform qualities but can be repeatedly divided and reunited at will to represent fine-grained differences in the numerical values of things. Also, they have a high value to weight ratio, which is useful if wealth is to be transported in pockets.

There has been little computational, as opposed to historical, work on the development of universal means of exchange in MAS: Marimon et al. (Marimon, McGrattan & Sargent, 1990) describes investigations of the conditions in which money emerges in an artificial economy of adaptive, classifier system (Holland, 1986; Holland, Holyoak, Nisbett & Thagard, 1986) agents, although the chosen domain ontology bears only a superficial resemblance to real economies.

4.2 The function and properties of money

Money, therefore, is like any other commodity except for a social convention that ensures it is the means of exchange in all transactions. The particular form of value, be it gold, silver, bronze, paper or virtual currency flows, is a secondary matter: it is function that counts. The function and properties of money are now examined in greater detail. Importantly, the majority of these functions and properties have exact analogues in a computational setting.

(a) Money is a universal use-value. Money overcomes the limitations of bartering, eradicating the requirement for a local coincidence of wants and commodities. It is a commodity that all find useful. Producers become willing to exchange for a representation of value which has the functional property of being able to buy the products of others' labour. One effect of the introduction of money, therefore, is to free up the flow of commodities and increase the connectivity between agents. In a developed money economy everything has a price. Money may be exchanged for any product of any labour.

(c) Money has a well-defined, global meaning. The exchange-value of commodities as represented by the money-commodity is expressed quantitatively and is compared to other quantities of value. Consequently, the meaning of money is globally determined in a society of numerate agents.

(d) Money constrains possible exchanges. A loaf of bread may cost 50p but will not normally be exchanged for 49.5p because of the prevailing social convention. An agent with money can enter into many possible exchanges, whereas an agent without money cannot. The globally determined value of commodities defines what is and what is not a legal exchange, and serves as a kind of economic "all-or-nothing" law that controls the flow of commodities.

(e) Money has comparatively low communication costs. Consider the following thought experiment: instead of money exchanges a host of "middle-men" exchange lengthy notes listing individuals with their surplus-products and needs in an attempt to co-ordinate great chains of exchange mediated by coincidences of wants - a kind of global "swap shop". Such notes will entail high communication costs, due to the high information content of the notes, and high administration costs, such as matching up lists with lists. In direct contrast, money, being a number, is easily represented and removes the need for middlemen and their costly communications.[2]

(f) Money has comparatively low storage costs. The quality of money does not change. It can be stored by adding up all the quantities into a bigger quantity - a larger denomination of note, for example. There need be no storage of many qualitatively different things, such as filing cabinets of "co-ordination notes" in the above example.

(g) Money requires simple operators. Money requires only the very simplest operators: addition, subtraction and numerical comparison. No sophisticated local machinery is required to mediate the transaction. Money is quickly and easily parsed.

(h) Money can be accumulated. Money, if it is metal, such as gold, does not perish. It can be stored indefinitely.

(i) Money encourages the distal connectivity of producers. The coincidence of geographical location, time and wants for exchange to occur in a barter economy is overcome with the introduction of the money-commodity. Money can mediate wants, be easily transported from place to place, and be stored for future use, unlike perishables.

(j) Money is a domain-independent representation. In an exchange, value is compared with value. The value of a commodity does not represent anything external to the economy, nor does it represent any thing within the economy: it is internally relational, specifying an ordering over the set of commodities, including labour time. The precise nature of the ordering and how it changes in relation to changes in the economy as a whole is addressed in economic theories of value, a subject area characterised by historical controversy. The observation that money is a domain-independent representation does not rely on a particular theory of value. Domain-independence means that it would not make any difference to the functional role of money if the specific kind of labours within society changed or if the external environment changed.

(k) Money is part of a co-ordination mechanism. Importantly, money introduces supply and demand dynamics that implement a distributed solution to a global co-ordination problem. The co-ordination problem is how private labour can be co-ordinated on a social scale so that individuals' needs are met. Without a co-ordinating mechanism the social system would break apart; for example, basic goods might not be produced in sufficient quantities, or non-use-values (commodities that are not in demand) might be produced indefinitely.

Consider the following simplified scenario. An increase in productivity in one branch of production, say egg production, entails that the same share of the total labour of society can now produce more eggs. Assuming that demand for eggs is fixed the end effect of the increase in productivity is to free labour currently employed in egg production to be employed elsewhere in other branches of the economy. The value of commodities and the operation of the market is the mechanism that mediates this adaptive change. The total labour of society is dynamically allocated and reallocated in definite proportions to reflect changes in production techniques and demand for products. "It is only through the 'value' of commodities that the working activity of separate, independent producers leads to the productive unity which is called a social economy, to the interconnections and mutual conditioning of the labour of individual members of society. Value is the transmission belt which transfers the working processes from one part of society to another, making that society a functioning whole" (Rubin, 1988). Currency flow reinforces social co-operation: for example, a particular agent will not be able to acquire a commodity without first expending labour that has sufficient value to other agents. The market mechanism of exchange-value, the social convention of money, and the local reasoning of autonomous economic agents serves to meet the basic requirements of economic organisation outlined at the beginning of section 4.1.

5 Currency flow in multi-agent systems

This is all well and good, but what are the implications of the analysis of the role of money in a simple commodity

[2] But as we often discover to our cost, in some real and therefore less idealised markets, such as the housing market, chains of exchange and 'middle-men' do indeed occur.

economy for the design of adaptive multi-agent systems? In this section the particular form of value in economic systems is examined and compared to existing reinforcement learning algorithms, followed by a sketch of how currency flow could solve the problems of co-ordination in AMAS. Finally, a design hypothesis for AMAS co-ordination is proposed.

5.1 A universal, quantitative representation of value

All adaptive systems conform to the abstract schema of a selective system (Cziko, 1995), and all selective systems support concepts of value or utility (Pepper, 1958; Wright, 1997). A selective system has three components: (i) a trial generator, which is any mechanism that generates a variety of functions to produce outputs for particular inputs, (ii) an evaluator, which is a mechanism that evaluates the results of using particular functions to generate trials, where evaluation occurs through comparison to a norm, and (iii) a process of selection, which retains those functions associated with "good" evaluations for future use, while discarding others. Selective systems implement the well-known generate, test, and select cycle. Specific examples of selective systems improve their behaviour over time (cf. Darwinian evolution, genetic algorithms, classifier systems, neural networks, and adaptive multi-agent systems). In the abstract, economic systems are selective systems: the trials are the various concrete labours that produce commodities, the evaluation mechanisms are the various needs and demands of individual consumers, and selection occurs through the buying and selling of commodities. In an ideal market, what is produced matches what is required given available resources. Money mirrors the flow of commodities, reinforcing those productive activities that meet the demands of consumers. Human economic systems are an existence proof that exchanging numerical quantities can regulate complex processing systems. Information-theoretic analogues of some of the properties of currency flow identified in section 4.2 may be useful for co-ordinating adaptive, largely parallel information processing systems composed of autonomous agents (e.g., multiple instrumentality, semantic determinacy, low communication and storage costs, simple operators, domain-independence and the imposition of local constraints through the representation of global constraints). In fact, work in artificial intelligence uses economic ideas for resource allocation problems (Wellman, 1995), including allocation of processing time, and reasoning about plans (Doyle, 1994).

5.2 Generalised reinforcement learning

Reinforcement learning (RL) algorithms are selective systems as defined above (see (Kaebling, Littman & Moore, 1995) for a review). RL is a type of trial and error learning, and holds out the promise of programming control programs for agents by reward and punishment without the need to specify how a task is to be achieved. The main design problem to be solved in reinforcement learning is the credit assignment problem, which is the problem of "properly assigning credit or blame for overall outcomes to each of the learning system's internal decisions that contributed to those outcomes" (R. S. Sutton, quoted in (Cichosz, 1994)). More precisely, RL involves learning functions defined on the state and action space of a task, driven by a real-valued reinforcement signal. The details of how this is achieved depend on the particular function representation used. Examples of RL algorithms are Q-learning (Watkins & Dayan, 1992), classifier systems (Holland, 1975; Holland, Holyoak, Nisbett & Thagard, 1986; Wilson, 1995), and W-learning (Humphreys, 1996). Marvin Minsky's Snarc machine was an early reinforcement learner that encountered the credit-assignment problem (see section 7.6 of (Minsky, 1987)).

RL algorithms use a quantitative representation of value, the reinforcement signal, to select those behaviour-producing components that satisfy conditions of reward over and above those components that do not. Behaviour-producing components that have received high reward will be more likely to dispositionally determine the behaviour of the system in the future than those components with lower reward. For example, the bucket-brigade algorithm used in early classifier systems was inspired by an economic metaphor, in which system rules are agents consuming and producing internal messages (commodities) who each possess a certain amount of value (money) which they exchange for messages at a global blackboard (the market). Most RL algorithms are composed of rules. (Shoham & Tennenholtz, 1994) discuss a generalisation of RL to MAS called co-learning. Co-learning involves individual agents learning in an social environment that includes other agents. Co-learning agents must adapt to each other. (Kittock, 1995) describes some computational experiments on the emergence of social conventions through co-learning. Work of this kind is beginning to explore how MAS can adapt by reinforcement signals. The use of a universally recognised, domain-independent, quantitative representation of value is common to RL algorithms, co-learning, and economic adaptation via currency flow. However, the latter may require MAS with substantially

more sophisticated agents than those used currently. The theoretical relations at the information processing level of abstraction between reinforcement and payment for goods is an issue that can be fruitfully investigated by MAS research.

5.3 The ability to buy processing power

In economic systems and reinforcement learners, possession of "money" by an "agent" is a dispositional ability to buy processing power (Wright, 1996b). For example, a producer who makes a profit will have more money to employ more people (to buy processing power directly) and more raw materials (to buy the results of prior processing). Whether a thing is purchased or a person is purchased for a certain period of the day, an amount of labour power has been assigned to the purchaser. That the labour power has already been expended and is in the form of a commodity, or will be expended and is in the form of a commodity-maker, is a secondary matter. In both cases, processing resources have been bought. Individual profits and losses regulate this ability to commandeer and allocate social resources. Similarly, a rule in a classifier system uses its accumulated value to bid against other rules for messages in the "marketplace". Rules with high value are more likely to outbid rules of low value, process the message, and dispositionally determine the behaviour of the system. The bucket-brigade adaptively alters the ability of rules to buy processing power. The same holds for the weights of policy functions in Q-learning.

One of the most important scarce resources in a MAS is the agents themselves. The total processing power of the MAS is limited, where processing power is ability to do work. Similarly, Marx, drawing on the classical tradition in economics, emphasised labour-power as a finite resource in economic systems, developing the labour theory of value based on this conception. Labour-power is also the ability to do work. Whether it is computational agents performing abstract operations, or real people performing concrete operations, a transformation is taking place that can be called work.

Adaptive MAS must search for solutions to, perhaps continuously changing, global constraints. Therefore, there needs to be an ordering over the various agents of the adaptive system: some agents will perform more useful work than others with respect to certain constraints. The computational resources of the system should be concentrated on useful agents, be it in terms of giving them greater social power or allowing them access to more social products. In other words, useful work within a society (or useful processing within a mind) should be reinforced. The design principle of a quantitative representation of value that functions as an ability to buy processing power can integrate processing (useful computational work) and resources (limited computational power) with relatively low communication costs. Agents with more money can employ other agents, buy the products of other agents' work, and have greater control over system behaviour. Given these abstract and general considerations it is possible to sketch how currency flow could serve as a basis for co-ordination in adaptive multi-agent systems.

5.4 Specifying global constraints

Economic systems suggest a natural way to specify the global constraints of an AMAS. In simple commodity economies it is the wants of consumers that determines what is and what is not a use-value. In just so happens that in real economies consumers are normally also producers, but in artificial AMAS the functions can be separated and assigned to different agents. A set of consumer agents that function as the sole sources of payment can define the goals of the system. Producer agents must satisfy consumers' wants if they are to receive value for their work. It is feedback from consumers to antecedent producers in the form of payment that selects those productive behaviours that satisfy the global goals of the system, much as conditions for reward select adaptive policies in RL algorithms. For example, an AMAS may be designed to find plans for successful operation in a microworld domain, such as blocks-world. A set of consumer agents can be defined whose various needs are information items declaring that the system has achieved certain objectives, such as stacking a tower of blocks or building certain shapes and so forth. These information items are analogous to desired commodities in economic systems: they are the use-values of the system. A set of producer agents may then attempt to produce the required information items by performing work in the domain, that is produce information items interpretable as actions by a scheduler. Only those agents or group of agents that produce the correct set of actions and corresponding results receive money from the set of consumers. Partial solutions may receive partial payment allowing hill-climbing and iterative trial and error search. Baum (Baum, 1996) describes the "Hayek machine" that learns to solve blocks world planning problems using a free market of interacting agents and a simplified price mechanism. Weiss (Weiss, 1995) describes the "Dissolution and Formation of Groups" algorithm that solves block world

problems using a collection of agents that learn through reinforcement and form into co-operative groups with "leaders". The Contract Net Protocol (d'Inverno & Luck, 1996; Smith, 1980; Smith & Davis, 1981) has, for many years in the field of DAI, also embodied some of these economics-flavoured ideas. In a contract net, a manager agent broadcasts a task announcement message, and receives bids from contractor agents. The manager evaluates the bids, selects among them, and allocates the task, or part of it, to the best bidder.

5.5 Dynamic control relations

As stated, an AMAS may need to alter the control relations between agents in order to meet global goals. A relation of control exists between agent A and agent B if A can determine, or dispositionally determine, B's processing. For example, A may be able to command B to perform a particular task, or A may be only able to request that B perform a task in particular circumstances, and so forth. In human societies there is a wide variety of relations of control, some more benign than others. Autonomous agents will often have objectives that conflict with other autonomous agents. One way for agents to overcome conflicts of interest is through negotiation, a process by which a group of agents communicate with one another to arrive at a mutually acceptable course of action. For example, when a conflict is encountered the agents involved may generate proposals for joint commitments with associated explanations. The mooted proposals may then be evaluated, and various counter-proposals or compromises suggested. The Socratic dialogue continues until agreement is reached (Parsons & Jennings, 1996). However, this may be locally rational but globally irrational with respect to the overall goals of the social system.

In order that local negotiations can meet global requirements there is need for local information, referring to those requirements, that can form a basis for controlling the negotiations. Without such information agents could negotiate commitments that led to globally incoherent behaviour or that required too many resources (i.e., the construction of unrealisable social plans). In human societies many negotiations occur within the context of financial costs. For example, much institutional behaviour consists of negotiating compromises constrained by available funding. The local possession of value limits the formation of commitments, which are essentially about resources (Bond, 1990; Gerson, 1976). By giving access to additional resources, commitments thus become valuable resource in themselves (Aube & Senteni, 1996a; Aube & Senteni, 1996b). However, local possession of value can allow in turn the formation of new commitments. For example, a new injection of funding can release prior constraints on planning: planners may now have sufficient power to employ other agents to do their bidding or buy the resources needed to complete their plans. Money, as the ability to buy processing power, is an ability to form control relations; and *the flow of money adaptively allocates and reallocates constraints on local commitment formation.*[3] Again, one reason for this rests on the fact that commitments themselves constitute a special kind of resource, and that money embodies the value that is computed for these resources through social transactions. It is the requirement for global problem solving that necessitates the imposition of limits on local problem solvers: Hobbes chairs the Socratic dialogue. "Participation in any situation, therefore, is simultaneously constraining, in that people must make contributions to it, and be bound by its limitations, and yet enriching, in that participation provides resources and opportunities otherwise unavailable" (Gerson, 1976). Social agents commit to a social convention of money that simultaneously constrains and enriches possible local outcomes.

5.6 Dynamic reallocation of labour

An adaptive multi-agent system may need to reallocate agents to different tasks in order to meet global goals and maintain coherent behaviour. One possible solution is a global controller that has a wider picture of the whole system and directs the activities of others; however, keeping the agent informed could entail high communication costs, create a communication bottleneck, and render the other agents unusable if the controller failed (Jennings, 1996). The alternative is to distribute data and control, and economic systems suggest at least two possible mechanisms. A system composed of adaptive agents that attempt to maximise personal utility will exhibit distributed reorganisation of labour. Adaptive utility maximisers will search for rewarding tasks, allocating and reallocating themselves to different parts of the developing solution. For example, if a system constraint changes, such as a consumer agent requesting a qualitatively different result, then the agents that previously serviced the consumer will search for new forms of co-operation in order to produce the new result and regain gainful employment (c.f. rule discovery of rewarding areas of the pay-off landscape in classifier

[3] Compare (Bond, 1990; Gerson, 1976) where money is viewed as just another kind of resource.

systems). In addition, a system that allows agents to sell their processing power to employer agents will exhibit organisational control, which is a "centralised" reorganisation of labour. For example, sufficiently wealthy employers may direct and redirect the processing of large groups of agents, perhaps at the expense of relatively high communication costs within the organisation. In both cases, however, it is money that forms the basis of the allocation of labour, either as a universal want or an ability to buy processing power. Note also that areas of the search space may be redundantly assigned to multiple agents, much as competition occurs within branches of production in real economies.

5.7 The currency flow hypothesis

Given these theoretical considerations and an analysis of some examples of existing systems, the following design hypothesis is proposed:

> *The currency flow hypothesis for adaptive multi-agent systems*: Currency flow, or the circulation of value, is a common feature of adaptive multi-agent systems. Value serves as a basis for co-ordination; for example, it integrates computational resources and processing by constraining the formation of local commitments. Circulation of value involves (i) altering the dispositional ability of agents to gain access to limited processing resources, via (ii) exchanges of a quantitative, domain-independent representation of value that mirrors the flow of agent products. The possession of value by an agent is an ability to buy processing power.

The design hypothesis is a hypothesis because it is a statement about designs that can be falsified. It states something about the functional organisation of AMAS at the level of information processing. If the MAS research community discovers designs that meet the requirements for AMAS but do not use a currency flow mechanism then the hypothesis is falsified: the design feature is not common to that set of requirements. It is more likely, however, that the hypothesis in its current form is too general and imprecise. Future research may show that currency flow cannot meet all possible requirements for adaptive MAS behaviour, or that currency flow is necessary but not sufficient, or it is simply one of a range of possible alernatives, or it works for only certain types of constituent agents, and so forth. Therefore, the hypothesis serves as a guide, pointing towards perhaps fruitful areas of AMAS design-space based on an analysis of an existing, naturally occurring AMAS.

For a MAS to use currency flow mechanisms the constituent agents will need a minimal set of capabilities. A first pass requirements analysis suggests that minimally economic agents will need to be able to form mutual plans with other agents, possess planning capabilities to construct and choose between alternative possible options, handle money, reason about costs, negotiate, and take and give requests and commands. Without these capabilities the economic system may fail to use currency properly or fail to find solutions to global requirements.

6 Some common objections

An objection to a quantitative representation of utility is that it necessarily entails a "loss of information" in order to reduce incommensurable quantities and qualities to a single, common utility measure. It is rightly claimed that many real-world problems are difficult or impossible to formulate in terms of maximising (or minimising) a single common measure (e.g., see (Logan & Sloman, 98)). However, in claiming it is necessary that AMASs employ currency flow (a single utility measure) it does not follow that they cannot also employ other representations of utility, for example in the individual reasoning of constituent agents and "non-commercial" exchanges of information. The fact that money exists and functions as described is good evidence that a single quantitative measure can perform a useful and important function in an adaptive multi-agent system. To date there are no examples of modern economies that function without money.

Furthermore, qualitative representations of utility imply the explicit representation of domain features. For example, specifying a qualitative ordering such as "substate A is more useful than B for determining processing in circumstances C for purpose P" (e.g., see (Sloman, 69)) would require domain-knowledge that may not be available or would be costly to deduce, particularly in a system composed of local reasoners without access to the global information important for determining local utility decisions.

A stronger argument would show why a quantitative representation of value is necessary. The argument would be in the form of a mathematical proof, not a design hypothesis. The question, therefore, remains open and is not yet sufficiently well stated.

Another objection is that the mind is goal-directed but the free-market anarchic and therefore an "economy of mind"

is an insufficient explanation for intelligent behaviour. Stating that a society of mind will require an economy of mind does not imply that currency flow is the only method of global co-ordination. (Wright, 97) proposes that a mental "currency" is the mechanism by which reinforcement learning constrains the formation of higher cognitive functions such that they conform to adaptive limits. This is a restatement of some features of Freudian metapsychology.

7 Implications for Cognitive Science

> "... another family of agents might contrive to use a quantity that doesn't actually exist at all, but whose amount is simply 'computed'. I suspect that what we call the pleasure of success may be, in effect, the currency of some such scheme."

M. Minsky, The Society of Mind, p. 284 of (Minsky, 1987).

If the society of mind requires an economy of mind and the information processing level of the brain is organised in such a manner, then we would expect some evidence of currency flow in our mental flora and fauna. Wright (Wright, 1996; Wright, 97) presents a circulation of value theory of achievement pleasure and failure unpleasure that explains the valenced component of some emotional states. Very briefly, the monitoring of virtual currency flows performing credit-assignment can account for some forms of mental pleasure and unpleasure. The theory is related to Freud's concept of "psychical energy" or "libido"; however, the circulation of value sheds the connotations of vitalism but clarifies and extends the functionality of libido. This work builds on previous work with Aaron Sloman and Luc Beaudoin on cognitive modelling of the emotions (Sloman, 78; Sloman & Croucher, 81; Beaudoin, 94; Wright, Sloman & Beaudoin, 1996; Sloman, Beaudoin & Wright, 1994). It is a recurring assertion that there is a relative neglect of motivation and emotion in cognitive science. For example, Simon's seminal paper (Simon, 1967) was an attempt to answer Neisser's criticisms that information processing theories of mind cannot account for feelings. More recently, (Newell, 1990) lists motivation and emotion as missing elements that need to be included in more comprehensive information processing theories of mind. (Shoham, 1996) argues that AI can and should benefit from economic ideas, for instance modelling the cost and value of information. If economic ideas are applicable to artificial intelligence then they should also be applicable to natural intelligence and therefore be of relevance to cognitive science. The concepts of value, currency flow, and ability to buy processing power are a step toward this.

8 Implications for Economics

Economics studies past and present economic systems. When analysing current economic organisation there is an implicit assumption that free-market organisation is either arguably or provably the best way to meet important global requirements such as efficiency and democracy. There is very little comparative exploration of possible economic systems. Economics, unlike AI, does not attempt to create new kinds of systems and does not make extensive use of computational explorations. One reason for this lack is the difficulty of reasoning about economic systems, which becomes extreme if the economic systems are hypothetical. If the currency flow hypothesis is correct then AMAS researchers will begin to explore varieties of designs for economic systems, albeit satisfying requirements that are very different from the requirements for human economic organisation. The convergence of ideas from AI and economics could result in a new branch of design-based economics that compares how different natural and artificial economic organisations meet various social requirements. Defining what those requirements should be for human social organisation is arguably not the subject matter of economics.

9 Conclusion

A hypothesis was proposed stating that a currency flow mechanism is likely to be a common feature of adaptive multi-agent systems: "a society of mind will require an economy of mind". Currency flow is part of a co-ordination mechanism that adaptively allocates and reallocates the ability of constituent agents to form local commitments. The social convention of money integrates resources and processing by functioning as an ability to buy processing power.

Acknowledgements

Thanks to Michel Aube, Aaron Sloman and Tim Kovacs for comments and criticism on an earlier version of this paper.

References

Aube, M. & Senteni, A. (1996a). Emotions as commitments operators: a foundation for control structure in multi-agent systems. In *Proceedings of the Seventh European Workshop on Modelling Autonomous Agents in a Multi-Agents World*, MAAMAW '96, Lecture Notes in Artificial Intelligence. Springer-Verlag.

Aube, M. & Senteni, A. (1996b). What are emotions for? Commitments management and regulation within animals/animats encounters. In Maes, P., Mataric, M., Meyer, J.-A., Pollack, J., & Wilson, S. W. (Eds.), *From Animals to Animats IV, Proceedings of the Fourth International Conference on the Simulation of Adaptive Behavior*, pages 264-271. The MIT Press.

Baum, E. B. (1996). Toward a model of mind as a laissez-faire economy of idiots. In *Proceedings of the Thirteenth International Conference on Machine Learning.*

Beaudoin, L. P. (1994). Goal processing in autonomous agents. PhD thesis, School of Computer Science, The University of Birmingham.

Boardman, J., Griffin, J., & Murray, O. (1993). *The Oxford History of the Classical World.* Oxford: Oxford University Press.

Bond, A. H. (1990). A computational model for organization of cooperating intelligent agents. In *Proc. of the Conference on Office information Systems*, pages 21-30. Cambridge, MA.

Cichosz, P. (1994). Reinforcement learning algorithms based on the methods of temporal differences. Master's thesis, Institute of Computer Science, Warsaw University of Technology.

Cziko, G. (1995). *Without Miracles, universal selection and the second Darwinian revolution.* Cambridge, Massachusetts: The MIT Press.

d'Inverno, M. & Luck, M. (1996). Formalizing the contract net protocol as a goal directed system. In de Velde, W. V. & Perram, J. W. (Eds.), *Agents Breaking Away, Proceedings of the 7th European Workshop on MAAMAW*, Lecture Notes on Artificial Intelligence, No. 1308, pages 72-85, Berlin. Springer.

Doyle, J. (1994). A reasoning economy for planning and replanning. In Technical papers of the ARPA Planning Initiative Workshop.

Gerson, E. M. (1976). On "quality of life". *American Sociological Review*, 41:793-806.

Holland, J. H. (1975). Adaption in natural and artificial systems. The MIT Press. Holland, J. H. (1986). Escaping brittleness: the possibilities of general-purpose learning algorithms applied to parallel rule-based systems. In Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.), *Machine learning, an artificial intelligence approach.* Los Altos, California: Morgan Kaufmann.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: processes of inference, learning and discovery.* The MIT Press.

Humphreys, M. (1996). Action selection methods using reinforcement learning. In Maes, P., Mataric, M., Meyer, J.-A., Pollack, J., & Wilson, S. W. (Eds.), *From Animals to Animats IV, Proceedings of the Fourth International Conference on the Simulation of Adaptive Behavior.* The MIT Press.

Jennings, N. (1996). Coordination techniques for distributed artificial intelligence. In O'Hare, G. & Jennings, N. (Eds.), *Foundations of distributed artificial intelligence.* John Wiley & Sons.

Kaebling, L. P., Littman, M. L., & Moore, A. W. (1995). Reinforcement learning: a survey. *In Practice and Future of Autonomous Agents, volume 1.*

Kittock, J. E. (1995). Emergent conventions and the structure of multiagent systems. In Nadel, L. & Stein, D. (Eds.), *1993 Lectures in Complex Systems: the proceedings of the 1993 Complex Systems Summer School*, Santa Fe Institute Studies in the Sciences of Complexity Lecture Volume VI. Santa Fe Institute, Addison-Wesley Publishing Co.

Logan, B. & Sloman, A. (1998). Qualitative decision support using prioritised soft constraints. Technical report CSRP-98-14, University of Birmingham, School of Computer Science.

Marimon, R., McGrattan, E., & Sargent, T. J. (1990). Money as a medium of exchange in an economy with artificially intelligent agents. *Journal of Economic Dyanmics and Control*, (14):329-373.

Marx, K. (1970). *Capital, a critical analysis of capitalist production*, volume 1. Lawrence and Wishart. Originally published in 1887.

Minsky, M. L. (1987). *The Society of Mind*. London: William Heinemann Ltd.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Parsons, S. D. & Jennings, N. R. (1996). Negotiations through argumentation - a preliminary report. In *Proceedings of the Second International Conference on Multi-Agent Systems*.

Pepper, S. C. (1958). The Sources of Value. University of California Press. Rubin, I. I. (1988). *Essays on Marx's Theory of Value*. Montreal: Black Rose Books. Originally published 1928.

Shoham, Y. (1996). The open scientific borders of ai, and the case for economics. Available at URL http:// robotics.stanford.edu /users /shoham/ aiecon.html. Draft note written for *ACM/CRA/NSF worksop on Strategic Directions for Computing Research*, held at MIT in June 96.

Shoham, Y. & Tennenholtz, M. (1994). Co-learning and the evolution of social activity. Technical Report CS-TR-94-1511, Robotics Laboratory, Department of Computer Science, Stanford University.

Simon, H. A. (1967). Motivational and emotional controls of cognition. Reprinted in *Models of Thought*, Yale University Press, 29-38, 1979.

Sloman, A. (1969). How to derive "better" from "is". *American Philosophical Quarterly*, 6(1):43-52.

Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy*, Science and Models of Mind. Hassocks, Sussex: Harvester Press (and Humanities Press).

Sloman, A. & Croucher, M. (1981). Why robots will have emotions. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver.

Sloman, A., Beaudoin, L. P., & Wright, I. P. (1994). Computational modeling of motive-management processes. In Frijda, N. (Ed.), *Proceedings of the Conference of the International Society for Research in Emotions*, Cambridge. ISRE Publications.

Smith, R. G. (1980). The contract net protocol: high-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*, 29(12):1104-1113.

Smith, R. G. & Davis, R. (1981). Frameworks for cooperation in distributed problem solving. *IEEE Transactions on Systems, Man and Cybernetics*, 11(1):61-70.

Watkins, C. & Dayan, P. (1992). Technical note: Q-learning. In *Machine Learning 8*, pages 279-292.

Weiss, G. (1995). Distributed reinforcement learning. *Robotics and Autonomous Systems*, 15:135-142.

Wellman, M. (1995). Market-oriented programming: some early lessons. In Clearwater, S. (Ed.), *Market-Based Control: A Paradigm for Distributed Resource Allocation*. World Scientific.

Wilson, S. W. (1995). Classifier fitness based on accuracy. *Evolutionary Computation*, 3(2):149-185.

Wright, I. P. (1996). Reinforcement learning and animat emotions. In Maes, P., Mataric, M., Meyer, J.-A., Pollack, J., & Wilson, S. W. (Eds.), From Animals to Animats IV, *Proceedings of the Fourth International Conference on the Simulation of Adaptive Behavior*, pages 272-281. The MIT Press.

Wright, I. P. (1997). Emotional Agents. PhD Thesis, School of Computer Science, The University of Birmingham. (Available online at http: //www.cs.bham.ac.uk/ research/ cogaff/).

Wright, I. P. & Aube, M. (1997) The society of mind requires an economy of mind. Technical report CSRP-97-6, School of Computer Science, University of Birmingham.

Wright, I. P., Sloman, A., & Beaudoin, L. P. (1996). Towards a design based analysis of emotional episodes. *Philosophy Psychiatry and Psychology*, 3(2):101-137.