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 *Gedanken-experiments*, 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.

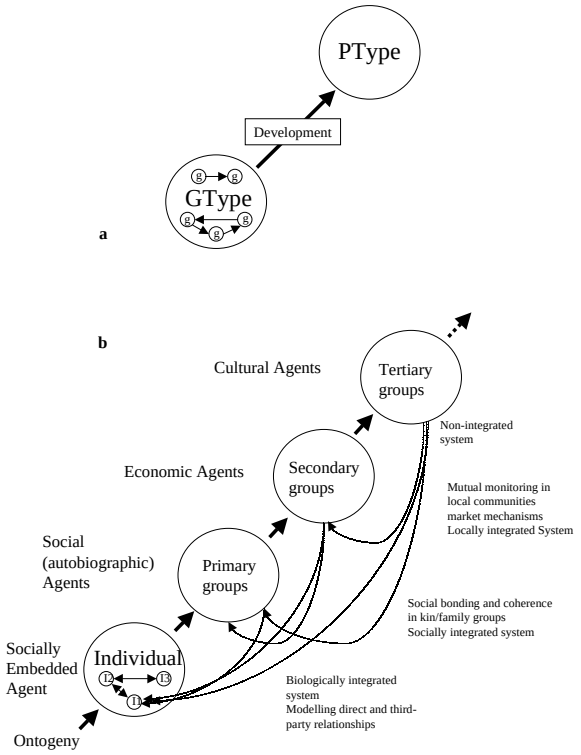

Figure 1: a) Emergence of behaviour in anonymous societies, b) emergence and feedback in individualised societies of socially embedded human agents on different levels of social organisation

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 lifelike 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.