Agent Cognitive Ability and Orders of Emergence

AISB 2008 Proceedings Volume 6

AISB '08

GLoRiClass

aisb ABERDEEN - 2008

UNIVERSITY OF ABERDEEN

AISB 2008 Convention
Communication, Interaction and Social Intelligence

1st-4th April 2008

University of Aberdeen

Volume 6:
Proceedings of the
AISB 2008 Symposium on Agent Cognitive Ability and Orders of Emergence

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

http://www.aisb.org.uk/convention/aisb08/

ISBN 1 902956 65 6

Contents

The AISB'08 Convention: Communication, Interaction and Social Intelligence

As the field of Artificial Intelligence matures, AI systems begin to take their place in human society as our helpers. Thus it becomes essential for AI systems to have sophisticated social abilities, to communicate and interact. Some systems support us in our activities, while others take on tasks on our behalf. For those systems directly supporting human activities, advances in human-computer interaction become crucial. The bottleneck in such systems is often not the ability to find and process information; the bottleneck is often the inability to have natural (human) communication between computer and user. Clearly such AI research can benefit greatly from interaction with other disciplines such as linguistics and psychology. For those systems to which we delegate tasks: they become our electronic counterparts, or agents, and they need to communicate with the delegates of other humans (or organisations) to complete their tasks. Thus research on the social abilities of agents becomes central, and to this end multi-agent systems have had to borrow concepts from human societies. This interdisciplinary work borrows results from areas such as sociology and legal systems. An exciting recent development is the use of AI techniques to support and shed new light on interactions in human social networks, thus supporting effective collaboration in human societies. The research then has come full circle: techniques which were inspired by human abilities, with the original aim of enhancing AI, are now being applied to enhance those human abilities themselves. All of this underscores the importance of communication, interaction and social intelligence in current Artificial Intelligence and Cognitive Science research.

In addition to providing a home for state-of-the-art research in specialist areas, the convention also aimed to provide a fertile ground for new collaborations to be forged between complementary areas. Furthermore the 2008 Convention encouraged contributions that were not directly related to the theme, notable examples being the symposia on "Swarm Intelligence" and "Computing and Philosophy".

The invited speakers were chosen to fit with the major themes being represented in the symposia, and also to give a cross-disciplinary flavour to the event; thus speakers with Cognitive Science interests were chosen, rather than those with purely Computer Science interests. Prof. Jon Oberlander represented the themes of affective language, and multimodal communication; Prof. Rosaria Conte represented the themes of social interaction in agent systems, including behaviour regulation and emergence; Prof. Justine Cassell represented the themes of multimodal communication and embodied agents; Prof. Luciano Floridi represented the philosophical themes, in particular the impact of society. In addition there were many renowned international speakers invited to the individual symposia and workshops. Finally the public lecture was chosen to fit the broad theme of the convention – addressing the challenges of developing AI systems that could take their place in human society (Prof. Aaron Sloman) and the possible implications for humanity (Prof. Luciano Floridi).

The organisers would like to thank the University of Aberdeen for supporting the event. Special thanks are also due to the volunteers from Aberdeen University who did substantial additional local organising: Graeme Ritchie, Judith Masthoff, Joey Lam, and the student volunteers. Our sincerest thanks also go out to the symposium chairs and committees, without whose hard work and careful cooperation there could have been no Convention. Finally, and by no means least, we would like to thank the authors of the contributed papers – we sincerely hope they get value from the event.

Frank Guerin & Wamberto Vasconcelos

The AISB'08 Symposium on Agent cognitive ability and orders of emergence

The concept of emergence has become widely used within the agent community. However, it continues to be vaguely defined and to stand in for different propositions about social generative mechanisms. To date the community has focused primarily on upward causation (consistent with its usage within complex systems theory and artificial life) (Sawyer, 2003). Relatively little attempt has been made to re-examine critically the concept within the context of human agency. Similarly, derivative concepts such as downward causation and 'immergence' (Castelfranchi, 1998) have only recently begun to be explored in the simulation of human social systems.

Gilbert has referred to a form of emergence which cannot be explained using the conventional bottom up notion and which implies that emergence involving agents with advanced cognitive ability may be qualitatively different from when it is absent. This 'second order' emergence occurs, he argues, when agents recognize emergent phenomena, such as societies, clubs, formal organizations, institutions, localities and so on, where the fact that you are a member or a non-member changes the rules of interaction between you and other agents. In a similar vein, Goldspink and Kay (2007) have argued for the need to at least distinguish between:

- Non-reflexive emergence: where the agents in the system under study are not self-aware, and
- Reflexive emergence: where the agents in the system under study are self-aware and linguistically capable.

They have also attempted to identify the effect of these two orders on system characteristics and dynamics.

There is a need to advance the debate about the nature and form of emergence associated with human social systems and therefore relevant to human to human and human to agent interaction. Specifically there is a need to identify linkages between current theories of cognitive developmental thresholds, including but not limited to the development of language, narrative ability, self-identity and theory of mind, and to examine the implications that these developmental stages may have in supporting qualitatively distinct orders of emergence in social systems.

References

Castelfranchi, C. (1998) Through the Minds of the Agents. *Journal of Artificial Societies and Social Simulation*, 1(1) 5.

Gilbert, N. (1995). Emergence in social simulation. In N. Gilbert & R. Conte (Eds.), *Artificial Societies: the computer simulation of social life* (pp. 144-156). London: UCL Press.

Gilbert, N. (2002) *Varieties of Emergence*. Paper presented at the Social Agents: Ecology, Exchange, and Evolution Conference Chicago.

Goldspink, C. & Kay, R. (2007) *Social Emergence: Distinguishing Reflexive and Non-reflexive Modes*. Paper presented at the AAAI Fall Symposium Washington.

Sawyer, K. R. (2003) Artificial Societies: Multiagent Systems and the Micro-macro Link in Sociological Theory. *Sociological Methods & Research*, 31: 38.

Chris Goldspink & Nigel Gilbert

Control over Emergence

Martin Helmhout[1] and **Henk Gazendam**[2] and **René Jorna** [3]

Abstract. This paper explains and demonstrates emergence of organisational behaviour as a social cognitive mechanism, i.e. ones own behaviour at the cognitive level is influenced by interaction with others at the social level.

Besides the importance of understanding how behaviour evolves, it is probably more crucial to control emergence or enforce desired behaviour. In our research we demonstrate this by implementing social constructs as regulators or stimuli of behaviour.

The paper discusses a social cognitive architecture ACT-RBot or in short RBot[4] which is based on ACT-R. RBot inherits the cognitive architecture (production system) of ACT-R but provides also a mechanism of social constructs as meta-productions that operates as a social (control) layer. The architecture is implemented in software agents who 'live' in a discrete event simulation environment allowing them to interact and exchange signs.

The combination of RBot and a simulation environment provides observation of behaviour between agents (at the social level), but also introspection of the experiences of the individual agent stored in memory of its cognitive architecture.

We describe two simulation experiments that demonstrate the working of the social cognitive architecture. The first experiment shows that emergence is present at the cognitive (individual) and social level (interaction). The second experiment adds social constructs and authority that allows for (more) control over emergence.

1 INTRODUCTION

The problem with emergence is that it is often difficult to find out what caused it to appear. Secondly, when finding out, it is probably even more difficult to control, coordinate (or prevent) emergence within predefined boundaries.

For the social scientist, emergence is often described at the social level or level of interaction. The cognitive scientist on the other hand tends to focus on the way patterns in the mind of the individual emerge. According to Castelfranchi [11], a significant theoretical problem exists in the field of social sciences. There is a lack of understanding or explanation of unconscious, unplanned forms of cooperation among intentional agents. Cognitive science–in particular Artificial Intelligence–can contribute to the explanation of social phenomena. Likewise, cognitive science needs social science in order to incorporate social factors, i.e. there is a shortage of good ideas and theories that address socio-cultural concepts/signs/symbols with social structures from a cognitive standpoint [44].

In this paper we describe how to connect both sciences (micro-macro link [1]) and in particular explain cognitive emergence caused by social emergence and the other way around. We will also make a distinction (see [24]) between uncontrolled (first order) or spontaneous social emergence and controlled (social) emergence. During the start of the former type of emergence agents are first self-aware and strive to the outcome of their own actions in their best self-interest, but in the later state of emergence, conditioning (unconscious learning [3, p. 95]) takes place and the agents behaviour becomes more automatic. In the case of controlled social emergence, an agent requires social capabilities as well (e.g. social contructs, language,...), i.e. the agent needs to perceive (receive) and produce social constructs or activate them in the presence of relevant stimuli. The social construct serves as a moderator that influences the behaviour of the agent indirectly.

The structure of the paper is as follows. In section 2 we elaborate about social constructivism and social constructs. Section 3 discusses the social cognitive architecture RBot and section 4 describes experiments that demonstrate emergence as well as control over emergence. Finally, section 5 ends the paper with a short discussion.

2 SOCIAL CONSTRUCTIVISM

We argue that first order emergence, which does not require social skills, requires cognitive skills and some form of self-awareness[5]. Although such (economic) agents can be very successful, they are unable to comprehend the needs of others and therefore often strive to local optima instead of the optimal solution for a group as a whole. Economic agents equipped with sophisticated utility functions do not describe any actual economic or any other behaviour of any individual or group of individuals. "...economic agents are not socially embedded in the sense that the behaviour of no individual is influenced by interaction with any other individual" [32, p. 394].

Socially embedded agents are agents who are capable of expressing social behaviour and are aware that they are not alone, but are also part of a group or society. In other words, the agent does not only live in his own world (internal representation) but also builds up relations with the outside world; with physical, social and cultural objects, agents and groups of agents.

Social constructivism as a social psychological theory attempts to explain the relation between society and the individual who is part of that society. Mead [31] analysed the human being as a complex individual who socially constructs the world, itself and other human individuals; social construction of the world is created by a process of interaction. Hacking defines social construction as follows:

> X[, as a social construct,] need not have existed, or need not be at all as it is. X, or X as it is at present, is not determined by the nature of things; it is not inevitable.... X was brought

[1] ACIS, Guildford, United Kingdom, email: martin@acis.nl
[2] University of Groningen, The Netherlands
[3] University of Groningen, The Netherlands
[4] http://act-rbot.sourceforge.net

[5] We assume that the agent is cognitive plausible, has memory and therefore is able to reflect on its own past actions.

into existence or shaped by social events, forces, history, all of which could well have been different. [26, pp. 6-7]

Hence, a social construction or social construct can be seen as an invention or artefact (cf. [25, 39]) constructed by interaction between members of a social group or interaction between groups. Products of social construction, such as institutions, gender and emotions are social constructs, created, disseminated, and agreed upon by social groups [40, p. 522].

2.1 Affordances, signs & social constructs

Organisational semiotics [41, 42] suggests to combine *affordances* and *signs* to bridge the gap between the social and the individual level of the agent.

> Affordances stress the interaction between a human agent and its environment based on behaviour patterns that have evolved over time in a community. Signs stress the social construction of knowledge expressed in sign structures... Stamper sees affordances as repertoires of behaviours and distinguishes physical affordances and social affordances. [22, pp. 7-8]

A (physical) *affordance* is a set of properties of the environment that makes possible or inhibits activity [23]. After many encounters with the environment, this can result in a *habit of action*, which is a commitment to act with a connected action program that governs the actual acting [19, 36]. From a semiotic point of view, one could say that a physical affordance becomes a *social affordance* as well, the moment the physical affordance is shared between agents in a community. The experience of the object (shared with others) is built up in the mind of the agent; the agent is socially situated through interaction and perception, which is a process of social construction of signs in the agent's mind. The resulting signs are organised as units of knowledge consisting of a representation of an affordance and its associated habit of action.

Social constructs are social affordances [28, 41] and can be seen as representations of cooperation and coordination, based on intertwined habits and mutual commitments that are often expressed in sign structures such as agreements, contracts and plans. A social construct [20, 28] is a relatively persistent socially shared unit of knowledge, reinforced in its existence by its frequent use. In organisations, social constructs take the form of, for instance shared stories, shared institutions (behaviour rule systems), shared designs, shared plans, and shared artefacts. These social constructs support habits of action aimed at cooperation and coordinated behaviour.

In order to use social constructs in formal simulation models, we have defined a (not limitied) set of properties of a social construct [27].

- *Attached norms or rules*: social constructs can contain (a collection of) norms or rules that guide action and prescribe appropriate behaviour in a certain context. Our daily encounters with social norms (and law) are evident, for instance, when we are driving with our car on the right side of the street, or being polite for letting the elderly sit in the bus, etc.
- *Written/unwritten (coded/sensory)*: a social construct can be formed and communicated by writing the attached rules and norms down on paper, or they are internalised in agents and with the help of interaction transferred (language or gestures) to others [29].
- *Life span*: every social construct has a starting time, an evolution, a monitoring and controlling period and a finishing time [28]. The life span of every social construct, be it a norm, an emotion or an organisation varies and depends on other properties connected to the social construct, e.g. referred objects or facts in the social construct change over time, a lack of reinforcement of the social construct or changes in enforcement costs (people enforce others to obey the norms attached to the social construct).
- *Roles and identification*: the agent is given a role or identification, e.g. employer, employee, to make clear the authority, control and rules applied to that role [28].
- *Authority, responsibility and control*: according to Fayol [16], authority can be seen as 'the right to give orders' and the expectation that they are followed. Control and power can assure that agents behave responsible; they can be part of a directly involved authoritarian party or an independent third party. Assigning authority to someone creates a responsibility for that person to give orders and control whether other agents take their responsibility in following the 'rules of the game'.
- *Inheritance or prerequisite of other social constructs*: a social construct can become part of a complex network of connections with other constructs (that are often the result of earlier agreements). For example, when preparing a sales contract, sales men refer to their conditions that are registered at the institution of commerce and the registered conditions inherit conditions from public law
- *Scenario*: there can be a more or less standardised process (scenario / script [38]) for establishing a social construct between agents. Scenarios are often put on paper in which a specific order of social constructs over time is written down. In communities, scenarios are often informal and expressed in rituals and transferred from generation to generation.
- *Context*: context as property is a debatable issue, however there are two interpretations of context. Context can be situated *outside* the agent, and—possible at the same time—situated *inside* the agent, i.e. context is represented as symbols in the 'real' world stored externally from the mind and also as symbols stored in the mind of the agent. The external context contains certain elements—so-called affordances—perceived by the agent to which it is sensitive or is triggered by. In contrast to Gibson [23], and similar to Stamper [41, 43], Vera and Simon [48, p. 41] state that affordances "are carefully and simply encoded internal representations of complex configurations of external objects, the encodings capturing the functional significance of the objects". We assume that there has to be sufficient coherence between an internally represented social construct and an element in the external environment in order to activate or trigger the social construct. According to Gazendam, Jorna, and Helmhout [21]:

  > The recognition of a situation in which a [social construct] must become active must not only depend on the recognition of physical affordances like other agents, objects and situations, but also on an ongoing monitoring in the agent's mind of the state of the social context in terms of invisible entities like positions of rights and obligations of himself and of other agents, agreements and appointments that must be held, and so on. (p. 2)

Social constructs with all their properties are internalised and established by a process of socialisation and communication. During their life-span, they are monitored by enacting agents through observation, mentoring, practise, and training [29]. Secondly, they create standards of appropriate behaviour and stabilisation, i.e. they create shared expectations of behaviour and when exercised, they are evaluated by others as well. Thirdly, when they are widely known and ac-

cepted as legitimate, they are often self-enforcing, and its associated psychological cost of rejection will rise, e.g. the individual or group involved can feel psychological discomfort whether or not others detect the rejection [30]. And fourthly, because social constructs are connected to roles and authority, they can create formal and informal social structures. Such a social structure exists out of a network or collection of social constructs. When a social structure is legitimate, legally acknowledged and officially registered, it is often referred to as an organisation or institution.

2.1.1 Emergence of social constructs

The general processes that underlie social construction are the creation of new social constructs, the evolution, the control and ending of those constructs over time. Emergence of social constructs is mainly concerned with creation and evolution and with a lesser extent with control and ending of emergence. For a first order of emergence this is not a problem because it concerns mainly creation and evolution. The agents lack a sense of social affordance; they are insensitive to social control and sometimes even worse, they cannot be stopped. On the other hand, second order emergence cannot take place without social construction. We argue that agents need to be self and socially aware in order to control or even stop emergence.

We distinguish three phases during the life span of a social construct: creation, evolution and ending of the social construct. The control function is a separate process that can be applied during the entire life-span, i.e. during creation, evolution and ending of a social construct.

The process of *creating* a social construct can occur in three ways [17]:

1. Based on adaptive emerging social behaviour; when two agents have a conflicting or mutual goal / interest. For instance people going out of the elevator while others only can go in when it is empty, or a group meeting in which people after a couple of times play their (habitual) role.
   (tacit agreement in a community)
2. The other is communicative action or an authoritative ritual in which with help of communication, e.g. speech acts, a social construct is formed; it starts with the agent that wants to propose, inform, or is requested to inform or propose a social construct to another agent.
   (agreement in a community)
3. The third formation process is external norm formation and a clear specification of roles, in which certain individuals are authorised to prescribe and enforce norms, thereby regulating the individuals they supervise [49].
   (creation on authority)

The first creation process is an implicit social exchange of signs and agents play their (habitual) role without being aware of creating order and a new social construct. The last two creation processes are explicit. The agents communicate or bargain openly about what they want. During the *evolution* of a social construct, agents practise the social construct and learn by interaction with the physical, social and cultural environment: a social construct is adopted, reinforced, adapted or rejected. Normally, the *ending* of a social construct is reached when the end of the (agreed) life span or situation is reached. However, during the process of evolving, the social construct is subject to changes; agents re-negotiate, renew the social construct or simply end their commitments. After the ending of a social construct, (cognitive) agents do not forget the success or failure or contents of a social construct; the agent can reuse the experience with a social construct as prior knowledge in a new negotiation process. Ending of a social construct can also happen when it is not frequently used, i.e. the agent does forget how to apply the social construct in a specific context and therefore cannot commit herself anymore. Therefore, a large amount of social constructs are written down and exist in documents (e.g. law). For those reasons, many people make use of experienced agents (brokers, attorney) who have enough skills concerning social constructs and domain-specific knowledge.

Thus, the processes of creation, evolution and ending depend on the support for the social construct [7], such as the dynamics of the environment (e.g. entering and exiting agents), regime change and the control process that monitors these processes.

2.1.2 Control over emergence

Control processes determine the actual creation, evolution and ending of a social construct. During creation there is somehow a form of control, be it formal or informal, e.g. there is a formal set of rules actively enforced by an institution, an informal ritual process is known by the agents or agents are influenced by social constructs from their personal social and cultural historical background. Control is necessary in order to prevent agents from breaching agreements, become corrupt or even become a threat to society.

Agents feel the need to compare and find out if others are still aware of the conditions and associated norms of the social construct to which they all mutually committed. Conte and Castelfranchi [13] state that in a group of addressees to which a norm applies, the norm should be respected by all its addressees; it is a necessary consequence of the *normative equity principle* defining that agents want their 'normative costs' to be no higher than those of other agents subject to the same norms. Agents stick to a norm as long as the advantages of following up a norm outweigh the advantages of not following the norm, i.e. the agent has to comply with the norm and if necessary defend the norm or give up the norm (when the costs of obedience are too high compared to punishment).

The properties of social constructs are determined during the process of creation and evolution (emergence of social constructs). During that process agents desire to reach an outcome that is beneficial for the agent itself and/or for the community as a whole. Therefore, agents collect information about other's preferences concerning social constructs; especially when dealing with unexplored territory. Berger and Calabrese [9] state: "... [] when strangers meet, their primary concern is one of uncertainty reduction or increasing predictability about the behaviour of both themselves and others in the interaction (p. 100)". Hence, the desire of every agent is to get some form of control or get information about who is in control (authority) and what the 'rules of the game' are. Therefore, the agent (if he desires) has to spend time and energy (cognitive load) in finding out the required knowledge about how other agents behave and interact with each other (code of conduct). Often many struggles concerning power, authority and trust relationships first need to evolve before (tacit) agreements have been reached.[6]

Social constructs that have evolved and are considered stable, still require a certain amount of investment to be maintained in a community; control of emergence does not only concern the creation but the maintenance of the desired social construct as well.

[6] In this paper we do not elaborate about power and trust issues, but they certainly require attention because they are closely attached to (control over) emergence of social constructs

Sometimes it becomes rather difficult to change certain systems. Consider the following radical change of a social construct; starting now, everyone in the world should drive on the left side of the road which increases standardisation in car production in favour of a 'greener' environment. Such a change concerned today is impossible due to the huge amount of switching costs (and sunk cost). Besides that, politics, power and the feeling of uncertainty will keep the original social construct in place. Strong habits will make sure that there will not emerge a new system replacing the old one.

Summarised, control over emergence is desired when social constructs are created the first time introduced and when social constructs need to be maintained over a certain life-span. In section 4, we will demonstrate control over emergence in a simulation experiment.

3 Social Cognitive Architecture

The previous section elaborated that social constructivism enables the agent to become aware of the world around him and how this world can be represented in the mind of the agent. We argue that a social agent should have a cognitive architecture, i.e. be cognitive plausible in order to handle *representations* such as signs, language, relations and social constructs. A cognitive plausible agent (its architecture) is not only based on a physical symbol system [33], cognitive mechanisms, goal-directed behaviour and learning capabilities, but is empirically tested as well [34, 2, 47]:

> [A *cognitive plausible agent*] is a goal-directed decision-maker who perceives, learns, communicates, and takes action in pursuit of its goals, based upon [a physical symbol system] that implements theories of cognition [and is supported by empirical evidence]. [47, p. 88]

In our research we have adopted existing theories of cognition about cognitive architectures and compared the two dominant architectures–SOAR [34] and ACT-R [5]–and decided to adopt the theory ACT-R.[7] ACT-R can be described by three or four levels of description (see [14, 15]) to predict the behaviour of the overall system, i.e. the implementation level as an approximation of the physical level (the sub-symbolic level), the functional level–the functionality of procedures interacting with declarative chunks, and the intentional level–the beliefs, goals, desires and intentions.

Most cognitive architectures focus on the individual and more specifically on the cognitive band ($10^{-1} - 10^{1}$ sec) and the rational band ($10^{2} - 10^{4}$ sec) but *not* much on the social band ($10^{5} - 10^{7}$ sec) [34, p. 122]. ACT-R, as thoroughly empirically tested architecture, also does not pay much attention to (social) interaction between individuals.

Therefore, a new social cognitive architecture (RBot) was designed completely from scratch. RBot follows closely the cognitive theory of ACT-R, but some changes/adjustments were necessary to make the cognitive agent social:

- *Multi-Agent System:* social cognitive agent based simulations require artificial task environments in order to construct various interaction scenarios. We implemented a (environment) server that has the purpose of providing an artificial 'physical' task *and* communication environment. It allows agents to perceive objects and other agents, and to exchange signals and signs. Secondly, an Agent Communication Language (ACL) is implemented that enables ACT-R agents to communicate with each other.
- *Social interaction:* for a cognitive agent to become social, it needs to create and maintain shared knowledge, be able to socially construct its environment and represent social structures (institutions) and habits of action in its mind. In other words, the cognitive agent needs to understand and be able to express itself at the social level (band) as well.
  At the cognitive band, productions fire on average at every 100 ms (see [5]) and the rational analysis takes place at the rational band. Apparently there is a gap ($10^{1} - 10^{6}$ sec) between the social band and rational/cognitive band that needs to be bridged. In order to bridge this gap, we have defined the social construct as a (social) representation in the mind of the agent, or as documents or artefacts in society. The social construct allows us to define social situations in which the agent has to obey or is permitted to follow certain rules that are present in a society or culture. We have extended ACT-R with a special module (see figure 1) containing social constructs (as chunks) that influence certain aspects of behaviour.
  In general, social constructs can be formed in many ways; created by internal explicit or implicit cognition, or perceived from the external environment. Once created, a relatively simple mechanism can describe how social constructs operate. A social construct can have norms attached to it that can respond to changes of social representations in memory that reflect changes in social situations in the environment. The *condition side* of a norm is triggered by a combination of general (social) concepts or more specific instances of concepts. The *action side* has one or more targets (procedures/goals/other norms/etcetera) connected to it. As soon as the norm is triggered, it influences those target chunks by demoting or promoting their activation levels thereby indirectly changing the response functions and the behaviour of the agent.
  A social construct is a social chunk (concept) of cognition that can be linked to other social constructs creating a complex semantic network of constructs and norms. The mechanism or architecture of social constructs resembles the subsumption architecture of Brooks [10]; it is a multi-level system in which the upper (normative) level is a (network of) social construct(s) that influences behaviour at the lower level(s)[8].

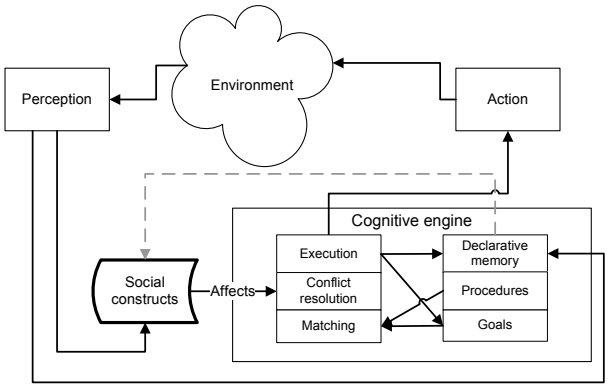

Figure 1. The extension of the cognitive engine with social constructs

- *Adjustment of rational and functional level:* we transformed ACT-R in order to be able to function at the social band, e.g. a slower de-

[7] We will not go into much detail about cognitive architectures here. For in-depth discussion, we refer to chapter 4 of 'The Social Cognitive Actor' [27]

[8] We actually made an attempt to combine ACT-R (hybrid agent: connectionism & symbolism (GOFAI)) with embodied cognition ('New AI')

cay of memory [46, p. 217]. Similar adjustments have been made in changing effort parameters of productions.

Much research still has to be done to make the architecture match its social requirements, however the current proposed mechanism will be a good start in making a cognitive architecture social. We will leave further implementation details behind us and start discuss the necessary parts of the social cognitive architecture that help to clarify the experiments demonstrated in section 4.

3.1 Base-level activation/decay & event discounting

Base level activation is probably *the* most important feature of ACT-R, i.e. it has been used in many environmental experiments [6] and has been the most successfully and frequently used part of the ACT-R theory [4]. Base-level activation B_i is an estimation of the log odds (of all presentations of a chunk in history) that a chunk will be used and is defined as:

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) + \beta \text{ Base-level Learning Equation}^{9} \quad (1)$$

- t_j represents the time-difference (t_{now} - $t_{presentation}$) that the chunk was represented in memory (created or retrieved),
- n the number of times a chunk is retrieved,
- d the decay rate,
- β the initial activation upon creation of the chunk.

The equation suggests that the more often a chunk is retrieved from memory (high-frequency), the more its base-level activation rises. On the other hand, the activation level of a chunk that is not retrieved at all can drop below an activation threshold level, whereby it becomes almost impossible to retrieve the chunk. Figure 2 exemplifies the retrieval of a chunk at t = 10 and t = 28. It shows a small (average) increase in activation over the specified period. As

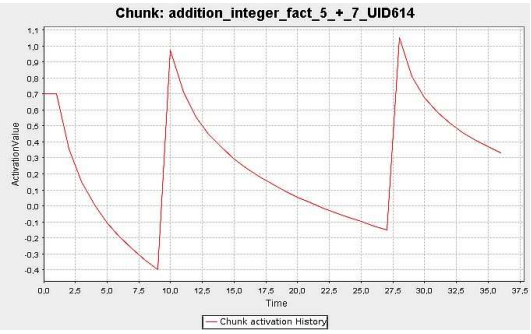

Figure 2. Example chunk, accessed at times 10 and 28.

regards the base-level learning equation, Anderson & Lebiere [5, p. 124] state that "odds of recall... should be power functions of delay, which is the common empirical result known as the Power Law of Forgetting [37]. Although it is less obvious, the Base-Level Learning Equation also predicts the Power Law of Learning [35]". Hence, the base-level learning equation enables the agent to learn and prefer chunks that are often needed for solving problems, and to neglect or forget chunks that are almost never needed.

[9] [5, p. 124]

For an agent to be socially situated, it needs to be able to learn facts and rules that emerge, but also to forget those when they are not reinforced or are replaced by other rules and facts in a community.

Decay functions are not only applied for chunks but also at the sub-symbolic level of procedures for events (success and failure). ACT-R has a mechanism that takes care of discounting past experiences by implementing an exponentially decaying function that is similar to the base-level learning equation. The equation for discounting successes and failures is:

$$Successes, Failures = \sum_{j=1}^{m,n} t_j^{-d} \text{ Success Discounting Equation}^{10} \quad (2)$$

- m number of successes,
- n number of failures,
- t_j time difference, now - occurrence_time of the success or failure,
- d decay rate.

The effect of the equation on the selection of productions is that there are more exploration possibilities caused by the decay of past successes and failures. For example, it is possible to give different decay rates for successes and failures, i.e. when the decay rate of failures is lower than that of successes, ACT-R tends to forget negative experiences more slowly than positive experiences.

Social constructs are reinforced by their frequent use. Hence, the abstract chunk with its base-level activation is an appropriate container for storing a social construct. The extension with the mechanism of the subsumption architecture combined with ACT-R triggers the social construct to become active when changes are taking place in certain areas of the memory (e.g. perception or communication).

The new model resulted in a social cognitive architecture called (ACT-)RBot which is part of an agent that operates in a Multi-Agent System giving space and time awareness to the agents and the ability to perceive and communicate with each other. In the following section we applied this architecture as the basis for our experiments.

4 Experiments

The purpose of the experiments is to show, as simple as possible, RBot operating in a multi-agent environment. Thereby demonstrating and finding out whether RBot agents can 'live' in a task environment and learn from interaction by cooperation, observation or other means. We have modelled a multi-agent environment of two agents in order to study the emergent behaviour of interaction between two individuals.

The general experiment is a case in which two agents have to pass each other as in a traffic experiment. They have to find a way out to pass each other several times without causing an accident. In the first experiment, there is no fixed traffic rule to drive on the left or the right side. In the second experiment, one of the agents acts as a controller (policeman) by enforcing the other to obey to the traffic rule he was given by higher authority. Hence, the policeman is present to control the way the social construct emerges in the mind of the other agent (given that the other agent respects the authority of the policeman) in case that agent wants to deviate from the common norm.

The basis configuration of both agents is equal; they are given the same parameters, procedures and declarative chunks, equal motivation values to solve goals and equal noise distribution functions. The

[10] [5, p. 141];[5, p. 265]

agents are identical in that sense that the simulation outcome is based on interaction and not on differences at the cognitive level of the agent. The agents learn in the following way; they evade right or left and then see if the evasion was successful or not. Based on this experience, the agent updates its productions and uses this experience in the next encounter.

The experiment showed two different emerging patterns. The first is an immediate lock-in; agents both initially choose the same strategy (e.g. driving on the right side), pass successfully and start to reinforce that particular strategy and leave no opportunity for the other strategy (driving on the left side) to emerge. The second pattern is caused when both agents at first do not select the same strategy, i.e. the agents are on a collision course. The second pattern is the most interesting, because agents have to decide based on experience which strategy is acceptable for both agents. By initially selecting the colliding strategy, the agents show a shifting behaviour from left to right and from right to left. After some collisions, the Boltzmann factor (noise) in the utility function gives the agent freedom of escaping from the hopping behaviour. This gives agents the opportunity to settle into the same successful strategy, both passing either left or right.

The interaction between two agents gives enough material to study phenomena and theories of different fields (e.g. management and organisation, social science, cognitive science, AI, to name a few). For analysis of organisational behaviour, we describe phenomena mainly at the individual level of description as a social construct (the behaviour of the individual agent), but we also describe behaviour at the social level by observing the interaction patterns between agents that emerge during the experiments. Hence, the experiments show behaviour of the individual and its internal cognitive properties and the emergence of behaviour of the collective as well (the behaviour caused by interacting or communicating (social constructs) between individuals).

4.1 Experiment 1: emergence of social constructs

This experiment demonstrates a first order of emergence in which we will explain how a (internalised) social construct emerges and evolves over time. The experiment will highlight that the agents seem to have a 'social agreement', but they actually adapt their own behaviour to the environment including the other agent, i.e. they do not intentionally influence or change each other's behaviour to come to an agreement. Because of its mutual adaptation, a social construct or norm emerges and creates regular stable behaviour in an otherwise chaotic system.

The purpose of the experiment is to show that a social construct can emerge not only by looking at the agents' collective outcome or rational level in the way they socially behave, but also what happens at the cognitive level and especially the sub-symbolic level of each agent.

In order to understand the behaviour of the agents, we want to look for behavioural patterns at two levels, one for the behaviour of the agent that is reflected by its internal history of memory and the other in the role of the external observer that spots the agents having a collision and ending up in an emerged pattern. With help of those patterns, the behaviour of escaping and reinforcement can be explained.

Figure 3 shows the results of the experiment (from an external observers point of view) that started with a collision. First, no particular strategy is chosen; the right move as well as the left move is equally preferred. However, approximately after time step 700, the agents prefer the 'Left' strategy based on interaction and experience built up in memory over time. Due to this interaction, they develop a preference in favour of the left move so strongly that the agents only choose the left strategy, i.e. the utility difference between right and left becomes significantly large that the agents are locked-in into the left passing strategy.

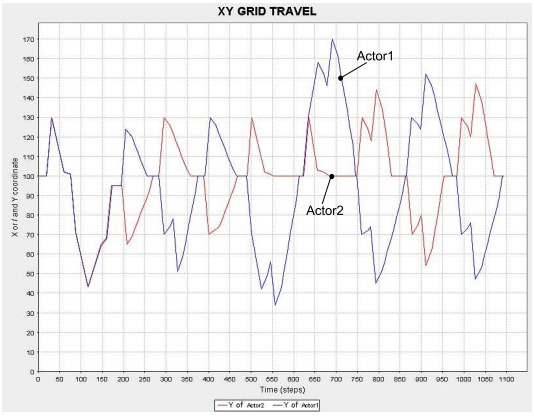

Figure 3. Y coordinates of both actors per time step[11].

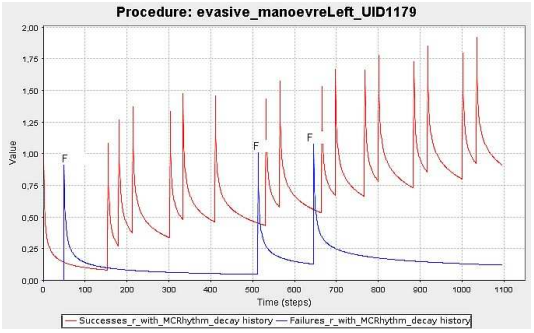

Figure 4. Agent 1: Successes, Failures Left

As an observer, if we only look at the social level and not at the cognitive level of the agents, we could conclude to see some start of "formalised" behaviour. However, after introspection of both agents (see figure 4 for an example of one agent), we observe at the sub-symbolic level of the production (strategy) 'evade left' that the agent reinforces this production, which results in a higher preference for this production than the production 'evade right'. We therefore conclude that the agent is aware of his own successes but does not care about the other agent's successes. In other words, there is not a formal agreement; the emergence exists as an evolved social construct (norm) in the head of both agents. Adaptation to their environment is purely based on their own individual experience (cognitive sub-symbolic learning).

4.2 Experiment 2: control over emergence

The first experiment, first order emergence, was mainly based on the interaction of RBot agents that make decisions based on their individual preferences. A cognitive agent solves a problem by defining a problem space and stating clear goals. Such an agent is however

[11] The graph shows the **y path** of two actors; the application is able to plot also the *x path*, but the y path is sufficient. The moment the paths are overlapping, the actors choose a different strategy and have conflicts because they follow the same y-path. When the paths are each others opposite, or mirrored, then the actors choose the same strategy and no conflicts arise.

not (socially) embedded in the environment, it responds not instantaneously to changes that occur in the (social) environment.

In this experiment we want to demonstrate the impact of social constructs, and particular how a social construct can be part of a coordination mechanism that allows an agent to get control over others' emerging behaviour(s). To get control over agents (car-drivers) and emergent behaviour, a police-officer can be assigned the task of controller (socially empowered by society) to correct behaviour of agents that do not behave according the society's norms or rules.

The experiment has the same initialisation as the previous experiment with the exception of the policeman. The policeman is given two social constructs: (SC1) The social construct of evading to the right side as stated by the government of, for instance, the Netherlands, and (SC2) communicate the social construct, SC1, of the preferred strategy to the other agent when he does not behave according to the norm (tries to collide by staying to long in front of the policeman). Hence, when a chance of collision starts to occur, the other agent needs to receive the social construct with the preferred strategy and be able to store it in memory and act accordingly.

The experiment is kept simple and only demonstrates the impact of a social norm on the behaviour of interacting agents. Therefore, the following assumptions have been made:

1. The agents are defined as being aware of their role, (1) Agent2 as policeman and (2) Agent1 obeying the policeman.
2. No negotiation takes place over the announcement of roles; the roles and relations are predefined by society.
3. Punishment is not included; however, it can be argued that by making the other agent aware of a certain social construct or norm is a correction of the behaviour. This can be regarded as a light form of punishment.

First we will have a look at the social or interaction level, see figure 5. We observe that when a collision is about to take place, Agent1 is immediately corrected in its behaviour and is forced by the policeman to choose the 'rightlanedriving' procedure. The control is so strong that after one encounter, the behaviour settles into a stable pattern. At the individual level of Agent1, see figure 6, we see a huge

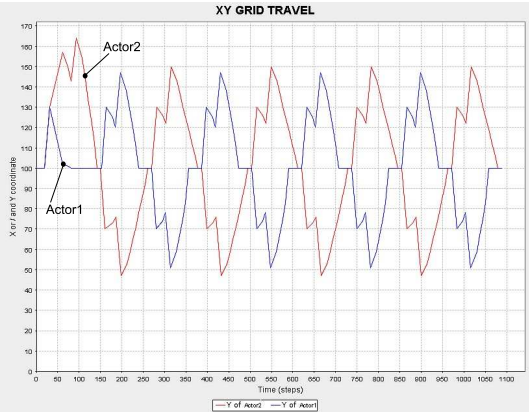

Figure 5. Y coordinates of both actors per timestep.

increase in utility in favour of driving on the right side. The correction being made is approximately a factor 10 and shows that Agent1 has become aware of the social construct 'rightlanedriving'. Besides that, the application software allows us to inspect the memory where we also observed that the social construct has been stored in memory of Agent1.

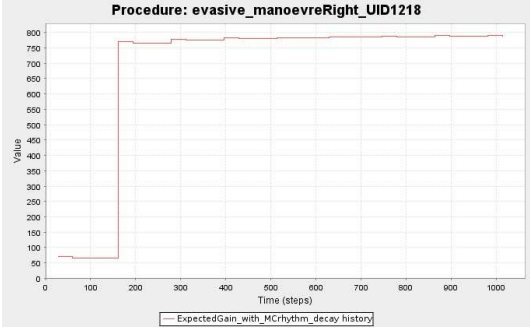

Figure 6. Agent 1: Utility Right

The experiment demonstrates clearly that the policeman is able to control the start of the emergence and keep control over the emergence by sending the social construct with the preferred strategy, be it by speech or by signalling his hand.

5 DISCUSSION

Emergence is around us in all kinds of varieties. However, to note that there are some properties that causes a system to emerge into a certain direction is not good enough. One has to look for a general explanation of (social) phenomena by studying not only the social or group level but the cognitive or individual level as well. Especially when studying second order of emergence in which individuals become aware of the emergence and the implications for others and themselves, then one *has* to look at the cognitive level of the agent and wonder if a differential equation still is powerful enough to give sufficient explanation.

We argue that especially in social situations an agent has to be able to built up representations in a cognitive plausible way, i.e. an agent should be able to express itself in a language and exchange representations with other agents. We have adopted the cognitive architecture ACT-R because it is empirically grounded and as a hybrid architecture allows for storing representations (symbolism) but connectionsm-like properties (activation) as well.

In the case of second order emergence, social constructs enable us to model social phenomena as a separate module (modularity of the mind [18]) of representations while leaving the underlying cognitive model of ACT-R as much as possible the same. The social construct mechanism described in this paper is conform the mechanism proposed by Bargh [8, p. 115] who also states that there are mental links between representations of motives and goals in memory and the representations of the social situations / constructs in which those motives have been frequently pursued in the past.

In cognitive psychology [45] and social psychology [12] empirical research has shown that there is a "dual process model", i.e. there is a distinction between implicit and explicit processes respectively automatic (reflexive) and controlled (reflective) processes. In this paper we did not put the emphasis on duality, but we can assume that social constructs or procedures, after they are first explicitly (or consciously) processed, become more of a habit or implicit/automatic process after emergence has taken place. The strengthening of the habit can be caused by the implicit association between situation and habit or by explicit reasoning (control) about the situation and the habit.

In case of control over emergence, it can be compared with government regulations that reduces uncertainty and enforces a system (society) to emerge in a way that is desired by society. However, many control mechanisms are based on semi-successful traditional

methods and are not based on research. For instance, we still do not understand how riots start to emerge; we can model group behaviour and take factors like weather condition, in-out group, relations, physical distance and so on to estimate how a riot emerges, but have no idea about the individual motives of taking part in a riot let alone how to model such an individual.

We present RBot as a general cognitive *and* social model that sheds light on the (intra)individual and the group level but not at the cost of too much complexity, lack of understanding and computational power. Besides that, its general application in the domains of social science and simulation, organisation studies, agent systems and cognitive science can hopefully revive the emergence of an interdisciplinary field and more cross-fertilisation in those areas.

REFERENCES

[1] J. C. Alexander and B. Giesen, 'From Reduction to Linkage: the Long View of the Micro-Macro Link', in *The Micro-Macro Link*, eds., J. C. Alexander, B. Giesen, R. Munch, and N. J. Smelser, 1–44, University of California Press, Berkeley, (1987).

[2] J. R. Anderson, *The adaptive character of thought*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1990.

[3] J. R. Anderson, *How Can the Human Mind Occur in the Physical Universe?*, Oxford University Press, 2007.

[4] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, 'An Integrated Theory of Mind', *Psychological Review*, **111**(4), 1036–1060, (2004).

[5] J. R. Anderson and C. Lebiere, *The atomic components of thought*, Erlbaum, Mahwah, NJ, 1998.

[6] J. R. Anderson and L. J. Schooler, 'Reflections of the environment in memory', *Psychological Review*, **2**, 396–408, (1991).

[7] R. Axelrod, 'An Evolutionary Approach to Norms', *American Political Science Review*, **80**(4), 1095–1111, (1986).

[8] J. A. Bargh, *Handbook of motivation and cognition*, volume 2, chapter Auto-motives: Preconscious determinants of social interaction, 93–130, New York: Guilford, 1990.

[9] R. C. Berger and R. J. Calabrese, 'Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication', *Human Communication Research*, **1**, 99–112, (1975).

[10] R. A. Brooks, 'A robust layered control system for a mobile robot', *IEEE Journal of Robotics and Automation*, **2**(1), 14–23, (1986).

[11] C. Castelfranchi, 'The theory of social functions: challenges for computational social science and multi-agent learning', *Cognitive Systems Research*, **2**, 5–38, (2001).

[12] *Dual-process theories in social psychology*, eds., S. Chaiken and Y. Trope, New York: Guilford Press, 1999.

[13] R. Conte and C. Castelfranchi, 'Understanding the functions of norms in social groups through simulation', in *Artificial Societies: The Computer Simulation of Social life*, eds., N. Gilbert and R. Conte, 252–267, UCL Press, London, (1995).

[14] D. C. Dennett, *Brainstorms: Philosophical essays on mind and psychology*, Harvester Press, Hassocks, Sussex, 1978.

[15] D. C. Dennett, *The Intentional Stance*, Bradford Books / MIT Press, Cambridge, MA, 1987.

[16] H. Fayol, 'Administration industrielle et generale', Technical report, Bulletin de la Sociéte de l'Industrie Minerale Dunod, Paris, (1916/1918).

[17] F. Flynn and J. Chatman, 'What's the norm here? Social categorization as a basis for group norm development', in *Research in groups and teams*, eds., J. Polzer, E. Mannix, and M. Neale, 135–160, JAI Press, Elsevier Science, London, (2003).

[18] J. A. Fodor, *The modularity of Mind*, Cambridge MA: MIT Press, 1983.

[19] H. W. M. Gazendam, *Information, Organisation and Technology: Studies in Organisational Semiotics*, chapter 1. Semiotics, Virtual Organizations and Information Systems, 1–48, Kluwer Academic Publishers, Boston, 2001.

[20] H. W. M. Gazendam, 'Models as coherent sign structures', in *Dynamics and change in organizations: Studies in organizational semiotics 3*, eds., H. W. M. Gazendam, R. J. Jorna, and R. S. Cijsouw, 183–213, Kluwer, Boston, (2003).

[21] H. W. M. Gazendam, R. J. Jorna, and J. M. Helmhout, 'Monitoring a social context based on social constructs', in *Interfacing Society, Technology and Organisations*, Campinas, Brasil, (2006). UNICAMP/IC.

[22] H. W. M. Gazendam and K. Liu, 'The evolution of organisational semiotics: A brief review of the contribution of Ronald Stamper', in *Studies in organisational semiotics*, eds., J. Filipe and K. Liu, Dordrecht, (2005). Kluwer Academic Publishers.

[23] J. J. Gibson, *The Ecological Approach to Visual Perception*, Houghton Mifflin, Boston, MA, 1979.

[24] N. Gilbert, 'Varieties of emergence', in *Social Agents: Ecology, Exchange, and Evolution Conference*, Chicago, (2002).

[25] N. Goodman and C. Z. Elgin, *Reconceptions in Philosophy and Other Arts and Sciences*, Routledge, London, 1988.

[26] I. Hacking, *The Social Construction of What?*, Harvard University Press, Cambridge, MA, 1999.

[27] J.M. Helmhout, *The Social Cognitive Actor: A multi-actor simulation of organisations*, Ph.D. dissertation, SOM research school: University of Groningen, 2006.

[28] K. Liu, *Semiotics in information systems engineering*, Cambridge Univerisity Press, Cambridge, England, 2000.

[29] J. G. March, M. Schulz, and X. Zhou, *The Dynamics of Rules: Change in written Organizational Codes*, Stanford University Press, Stanford, CA, 2000.

[30] R. H. McAdams, 'The Origin, Development, and Regulation of Norms', *Michigan Law Review*, **96**, 338–433, (1997).

[31] G. H. Mead, *Mind, Self and Society from the Perspective of a Social Behaviorist*, University of Chicago, Chicago, 1934.

[32] S. Moss, *Cognition and Multi-Agent Interaction*, chapter 15. Cognitive Science and Good Social Science, 393–400, MIT Press, Cambridge, MA, 2006.

[33] A. Newell, 'Physical symbol systems', *Cognitive Science*, **4**, 135–183, (1980).

[34] A. Newell, *Unified theories of cognition*, Harvard University Press, Cambridge, MA, 1990.

[35] A. Newell and P. S. Rosenbloom, *Cognitive skills and their acquisition*, chapter 1. Mechanisms of skill acquisition and the law of practice, 1–56, Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.

[36] C. S. Peirce, *Collected Papers*, volume I-VIII, Harvard University Press, Cambridge, MA, 1931.

[37] D. C. Rubin and A. E. Wenzel, 'One hundred years of forgetting: A quantitative description of retention', *Psychological Review*, **103**, 734–760, (1996).

[38] R. Schank and R. Abelson, *Scripts, Plans, Goals and Understanding*, Erlbaum Associates, Hillsdale, 1977.

[39] H. A. Simon, *The sciences of the artificial*, MIT Press, Cambridge, MA, 3rd edn., 1996.

[40] S. Sismondo, 'Some Social Constructions', *Social Studies of Science*, **23**(3), 515–553, (1993).

[41] R. K. Stamper, *Information in Business and Administrative Systems*, John Wiley & Sons, New York, 1973.

[42] R. K. Stamper, 'Organisational semiotics: Informatics without the computer?', in *Information, organisation and technology: Studies in organisational semiotics*, eds., K. Liu, R.J. Clarke, P. B. Andersen, and R. K. Stamper, Kluwer Academic Publishers, Boston, MA, (2001).

[43] R. K. Stamper, K. Liu, M. Hafkamp, and Y. Ades, 'Understanding Roles of Signs and Norms in Organisations', *Journal of Behaviour and Information Technology*, **19**(1), 15–27, (2000).

[44] R. Sun, 'Cognitive science meets multi-agent systems: a prolegomenon', *Cognitive Systems Research*, **14**(1), 5–28, (2001).

[45] R. Sun, *Duality of Mind*, Lawrence Erlbaum Associates, Mahwah, NJ, 2002.

[46] N. A. Taatgen, *Learning without Limits: From problem solving towards a unified theory of learning*, Ph.D. dissertation, University of Groningen: Faculty PPSW, Groningen, NL, 1999.

[47] H. Van den Broek, *On Agent Cooperation: The relevance of cognitive plausibility for multiagent simulation models of organizations*, Ph.D. dissertation, SOM Graduate School / Research Institute : Systems, Organization and Management, 2001.

[48] A. H. Vera and H. A. Simon, 'Situated Action: A Symbolic Interpretation', *Cognitive Science*, **17**(1), 7–48, (1993).

[49] F. A. Von Hayek, *Law, legislation, and liberty: The political order of a free people.*, volume 3, Routledge & Kegan Paul, London, 1979.

Cognitive architectures of agent systems and social mechanisms of emergence and immergence

Martin Neumann[1]

Abstract. This paper develops a framework of a theory for the emergence of social reality that turns out by relating three different models or agent architectures, respectively: a model of the emergence of social hierarchies, an architecture of a normative system and an architecture of the delegation of social control. This reflects a complex feedback loop of the emergence of an autonomous sphere of social positions and immergence of social norms and execution of social control. It is shown how social structure emerges from and recursively affects the agents' cognitive structure.

1 INTRODUCTION

The notion of emergence is well known in agent-based simulation as well as in Social Theory. The aim of this paper is to demonstrate that agent-based methodology can provide a tool for the formulation of an emergentist Social Theory. An emergentist theory of society is in need to express social micro and macro phenomena in terms of one another. It will be demonstrated that the cognitive architecture of software agents can provide a framework to investigate this question.

Within Social Theory emergentist theories promise to bridge the gap between the micro-macro distinction, or dichotomy of action versus structure [1 – 5]. Traditionally, so-called holistic and individualistic theories are opposing approaches [6, 7]. The former claim the existence of a social reality with laws and theories in their own right, not reducible to theories of lower level domains. This has been suspected as the error of *reification* [8]. The latter emphasises that this social reality does not exist. Every social phenomenon should be explained in terms of individual action. This has been suspected as the error of *voluntarism* [8]. The problem in analytically distinguishing micro- and macro-phenomena is that they do not appear separated but rather are inherent in one another [9, 10]. Theories of social emergence postulate the autonomy of social reality without denying that this reality is constituted by individual actors [5, 7, 11]. However, the ontological status of emergent social reality is often left unexplained [12].

Agent-Based simulation techniques promise to shed new light on this old problem by generating macro phenomena in the course of individual interaction. In fact, it is claimed that agent-based simulation provides a tool for studying emergent processes in society [13]. This promises to allow for an understanding how individual actors produce and are in the same time a product of social reality [14]. Thus, the methodology of agent-based modelling proposes an integrated view on the theoretical problem: While it is an individualistic assertion that actors produce social reality, it is a holistic position that actors are products of social reality. Hence, agent-based modelling techniques suggest to built-up a framework to understand micro- and macro-phenomena in terms of one another. This paper argues that agent-based simulation provides a tool to investigate the constitution of social reality not only by its generative capacity but also by the design of the agents' cognitive capacities: since agent-based modelling allows for an explicit modelling of the agents' cognitive capacities, it allows to study how the emergent level is constituted by the agents' cognitive design.[3]

The paper concentrates on *ontological* questions related to the constitution of social reality rather then on the epistemological question if and how emergence is possible (comp. [15]). Hints to the application of concepts of emergence in complexity research can be found in [16 – 19]. The relation to evolutionary processes is stressed in [20, 21]. The relation to psychology and philosophy of mind is investigated in [22 – 25].

The paper proceeds as follows: the first section deals shortly with the conceptual framework of emergence. The following sections relate models (or architectures) to processes in human societies. Since Society is not a physical object it will be asked how it can be realised by the agents' cognitive design. Two ontological dimensions are differentiated: First, the process of emergence of social positions is investigated. This is an evolutionary process of differentiation of social reality and individual actors. A sloppy phrase would be to denote this as externalisation of society.[4] Secondly, the reverse process of immergence of society by social norms is analysed. This is the causal effectiveness of social reality in the minds of individual actors. Finally, it is shown how both processes are recursively related by social control.

2 CONCEPTS OF EMERGENCE AND IMMERGENCE

According to Bedau [27], the basic assumption of emergentist theories is that reality is constituted out of a hierarchy of levels of reality, for which hold that:

[1] Institute of Philosophy, Bayreuth University, Universitätsstr. 30, 95447 Bayreuth, Germany. Email: martin.neumann@uni-bayreuth.de.

[3] Obviously, the agent architecture is not a realistic representation of human cognitive capacities. Insofar it provides only an investigation the *possibility* of the emergence of a new level of reality by up- and downward causation through cognition. However, evidence exist that human consciousness is processed by downward causal chains [22].

[4] The term is borrowed from activity theory [comp. 26]. In activity theory, however, the deployment of this term is not identical to the way it is used here.

a) emergent phenomena are somehow *constituted* by and generated from underlying processes and

b) emergent phenomena are somehow *autonomous* from underlying processes.

Bedau defines a phenomenon as emergent when a macrostate P of a system S with microdynamic D can be derived from D and S's external conditions but *only* by simulation [27]. This is a comparably weak definition of emergence [comp 15]. However, it is a tellingly characterisation of the processes at work in agent-based simulation. One of the features of Multi-Agent Systems is that they enable to built up patterns on the macro-level by local interaction of individual agents. The famous Schelling model [comp. 28] is one of the most prominent examples: Initially randomly distributed groups of agents produce patterns of segregation. The patterns of segregation are a newly generated emergent property of the social macro-level.

However, only recently attention has been paid to a related, but somewhat different problem: the way back from emergent macro-social properties generating effects on the micro-level. This complementary process has been denoted as *immergence* [29, 30], in Philosophy of Science also known as Downward Causation [31]. In a recent paper, Conte et al. [32] distinguish two main ways in which Downward Causation occurs in human and Multi-Agent Societies:

- a *simple loop*, in which the emergent effects produce new properties on the generating micro-level. Examples are dependence networks. Here the emergent macrostructure creates a distribution of negotiation power among individual agents at the generating micro-level [33]. This is a structural property of the network, independent of the individual consciousness of this structure.
- a *complex loop*, in which the emergent effect determines new properties on the micro-level by means of which the effect is reproduced again. Hence, a recursive interaction between both levels is established in such a complex feedback loop [32]. A particular interesting case occurs, when the emergent effect is recognised by the producing system. This has been denoted as 2nd order emergence [28, 34]. Stressing the crucial role of language, Goldspink and Kay [35] introduce the notion of reflexive emergence; an effect that as a matter of fact exist in Human Societies. For instance, people recognise norms and act (sometimes) accordingly.

This paper will investigate the cognitive capacities needed to generate reflexive emergence. In terms of Social Theory this would contribute to the above mentioned micro-macro problem. It would be a fundamental contribution to Social Theory to represent 2nd order emergence in Multi-Agent Systems.[5] In particular, representing agents' conscious awareness of an emergent social reality would be a building block to develop a theory of Sociality, which would be able to explain recursively micro and macro phenomena in terms of one another. Such an investigation demands an examination of the agents' cognitive capacities. However, concurrently a framework will be outlined of what social reality actually consist.

[5] For reasons of terminological simplicity, in the following the paper will simply refer to the term immergence to denote the process of downward causation.

3 EMERGENT SOCIETY

Translating the philosophical concept of emergence into material conditions of human societies needs to identify the emergent object. In case of human societies it is not so obvious what are emergent levels of social reality, since different levels are interweaved in individual action. Actors and society exist parallel. Even though this need not be an exhaustive enumeration, it will be proposed that it is possible to distinguish at least two forms in which emergent and immergent social reality can be identified. Emergence is specified as a process of differentiation of social positions and individual actors. Immergence is specified as the causal power of social reality on individual behaviour by norms.[6] Thereby society finds its way back into the minds of the individuals. First the process of emergence is investigated.

Following Peter Blau [36], the emergence of social structure will be conceptualised as the distribution of a population among social *positions*. This is because social structure "nearly always includes the concept that there are differences in social positions, and that there are social relations among these positions" [36, p. 27]. Undoubtedly, social positions influence people's social relations, but they have to be distinguished from mere interaction. For instance, at different times the same position can be inhabited by different people. Therefore positions gain an autonomous reality. Hence, by the establishment of social positions a differentiation between social reality and individual actors take place.

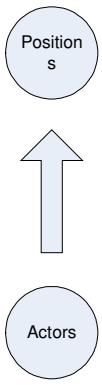

Figure 1. emergence of social positions

In fact, evidence exists that the emergence of social positions has been a concrete historical process which took place in space and time. Archaeological findings indicate that the process of differentiation between individual actors and social positions is the process of the emergence of social stratification, presumably located in the Palaeolithic period [38 – 42]. No archaeological indicators for social stratification can be found in earlier societies. In course of cultural evolution, however, "egalitarian principles of burials were violated when extraordinary items of gold started to be placed with certain individuals, presumably to

[6] There might exist more instances of social reality than social norms and positions. For instance, some sociological theories [e.g. 37] stress the crucial role of communication and thus language. Presumably language is a precondition for the generation of cognitive capacities that enable the emergence of social norms and positions. It is not denied that there might exist a hierarchy of several levels of social reality.

mark their social difference." [42, p. 112]. It is highly plausible that the process of social stratification goes hand in hand with the emergence of social positions. This is a process of the emergence of an autonomous element of social reality. The emergence of cultural symbols denote the emergence of a new form of reality, consisting of role differentiation and social stratification. However, since positions are no physical object, the emergence of positions involves some kind of transformation the cognitive structure of individual actors. In the following a model is analysed with regard to the question what cognitive capacities are needed for this innovation in the organisation of human relations.

agent models of social emergence

There exist archaeological models of the emergence of stratification. One – quite old, but in this respect still outstanding model will be considered briefly: the EOS model of the emergence of organised society.[7]

The target of the EOS model is to develop an agent-based model of a theory [38] of the growth of social complexity in the Upper Palaeolithic period of South–West Europe, that is 15 000 to 30 000 years ago [45]. In contrast to egalitarian societies, complexity is defined as containing centralised decision-making, ranking, role differentiation, and territoriality [46]. Hence, among other features, the evolution of social stratification and role differentiation is denoted by the notion of social complexity. The main features of the model are [45]:

a) a two-dimensional simulated environment providing clusters of resources that can be gathered by the agents. The resources have a specific regeneration cycle and complexity, which is defined as the number of agents necessary to acquire them.

b) a population of 32 to 50 agents. The agents are able to collect sensory data and move around in the environment. In particular, they form plans for resource acquisition and communicate about these plans. Agents need to gain resources.

To examine the question of how this model is capable to represent emergent sociality, it is of particular interest to investigate the agents' cognitive capacities: The working memory of the agents contains a resource model, where the agents keep their beliefs about the resources, and a *social model*, where an agent stores its beliefs about itself and other agents.

In course of the simulation the agents start without any knowledge of groups or other agents. They collect information about their environment and, if they are able to collect resources individually, they do so. If there are resources that need co-ordinated activity, then they develop plans for collective resource gathering and attempt to recruit others for the execution of the plan. Therefore they send out information about the resource and the others evaluate this information to decide whether or not to follow. Agents that are able to recruit others become group leaders. The agents, whose plans are selected, gain 'prestige'. This leads to a "semi-permanent leader-follower relationship" [45, p. 106]. This group structure becomes part of the social model of the involved agents. This process may be iterated, thus leading to a situation where a group leader together with its group members becomes a participant of another group. Hence, by iterations a social hierarchy is formed.

These hierarchies have the ability to persist, but may also break down after a while. This is affected by how easily agents decide to operate independent of their leader and how long they believe to be part of a group when they are not in contact with it [45, p. 214]. Hence, the EOS model allows the study of mechanisms of the emergence of social stratification out of egalitarian groups of agents.

Agent cognitive capacities

It thus has been concluded that the hierarchy is an "implicit property of the agents' social model" [47, p. 154]. However, regarded from the perspective of duality of structure and action, the way back from emergent structure (namely: hierarchies) into the agents' social model is a crucial property of the agents' cognitive structure. Namely, in this model social reality is expressed in terms of individual agents. The restriction of this model is not primarily that the hierarchy is a property of the agents' social model but that the agents' social model is restricted to knowledge about *individual actors*. It does not include the notion of social positions, where actors in such positions are responsible to form plans. In terms of George Herbert Mead [48], they lack the notion of the generalised other. This restriction of the agents' cognitive capacities leads to the result that the model is incapable of generating a social sphere of its own. It shall not be question whether the model is a correct representation of the social reality at around 20 000 BC.. However, the differentiation of social positions and individual actors is a qualitative switch in human prehistory that agent-based modelling technology has not represented so far. The cognitive capacities of the agents do not include a kind of *abstraction* process from individual leaders to the abstract notion of a leader and thus social positions. Eventually, the current New-Ties Project might provide new inside into such processes.[8]

4 IMMERGENT SOCIETY

The process of differentiation between actors and positions is a process of the emergence of a new level of social reality. It is represented both in material symbols of social status and in the mind of the actors, able to identify the symbolic meaning of the material items. Classical role theory [49 – 51] emphasises that positions are characterised by social *norms*.[9] Hence the emergence of positions includes the dual process social *immergence*. For instance, in the EOS model social hierarchies are a representation of the agents' social model. This is a complex feedback loop, in which the emergent effect is recognised and reproduced by the producing actors. This is a social norm: Norms are means to regulate individual behaviour in a way prescribed by society (for instance, the norm to respect

7 This model has been selected since more recent models such as those concerned with the decline of the Anasazi culture [43, 44] do not capture the process of social differentiation.

8 `http://www.cs.vu.nl/~gusz/newties/newties.html`

9 The immergent effect of norms is essential to establish social positions. However, norms may have existed before the emergence of social positions took place. Presumably, the emergence of norms went along with the emergence of language (compare also footnote 6). However, positions enable the establishment of a new kind of norms (such as money) for large and anonymous societies. Hume [52] denoted this as artificial virtues.

the authority of a group leader). Social norms are essential features in the coordination of populations of actors. Hence, by the means of social norms, social reality is causally effective in the minds of the individual actors [29]. In the following, a closer examination of agent-based models of norms is undertaken.

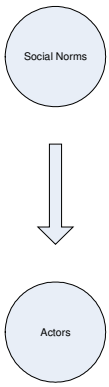

Figure 2. immergence of social norms

Agent models of social immergence

The current development of normative agent systems mostly follows the Belief-Desire-Intention Architecture [53], by extending this approach to a Belief-Obligations-Intention-Desire (BOID) Architecture [54]. Examples for such an approach are [55 – 60]. Obligations are introduced to constrain individual intentions and desires on the one hand, while preserving individual autonomy on the other [54]. Agents are able to violate normative obligations. This implies that agents dispose of the capacity of normative reasoning. A sophisticated agent design following a similar logic, however without an explicit notion of obligations can be found in [61].

In this paper, a closely related, but nevertheless somewhat different account is examined in more detail: Boella/van der Torre's 'An architecture of a normative system' [62]. This architecture is selected, since it relies on John Searle's theory of constructing social reality [63]. It is thus an approach to explicitly represent social reality in the agents' cognitive capacities. The primary technical terminus in Searle's theory are the so-called *counts-as conditionals*. Searle's theory of social reality distinguishes between brute and institutional facts. Institutional facts are build upon social norms. Two types of norms are distinguished: some norms regulate pre-existing forms of behaviour, while other norms create the possibility of that activity. For instance, playing chess is constituted by the rules of the game. Following Searle, the general form of a normative ontology is 'X counts-as Y in context C'. Boella/van der Torre's architecture intends to represent this theoretical approach in the agent's design.

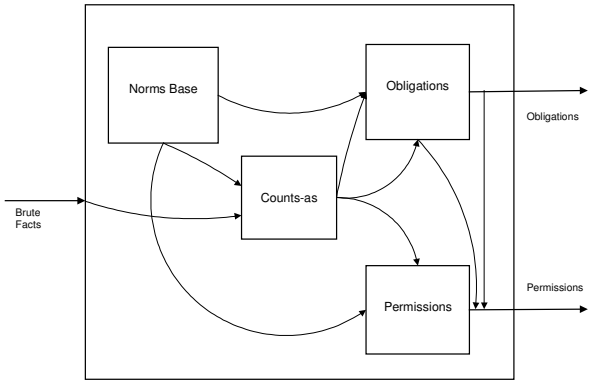

Figure 3. simplified Architecture of a Normative System

In the architecture the inputs of the normative system are brute facts. The outputs are obligations and permissions. The normative architecture has four components: counts-as conditionals, obligations, permissions and a norms base. Thus, if a proposition p is flowing over a permission channel it is interpreted as a permission P(p). The *norms base* has no input, it is assumed as fixed. It includes background knowledge of institutional constraints. The inputs of the *counts-as component* are brute facts, counts-as conditionals and institutional constraints. Counts-as conditionals and institutional constraints come from the norms base. Brute facts are external inputs. The outputs are brute and institutional facts. They are sent to the obligation and permission components. The input of the *obligation component* contains conditional obligations (coming from the norms base), a context (a mixed description of brute and institutional facts coming from the counts-as component) and constraints. Constraints are added to avoid inconsistencies. The outputs are obligations. The design of the *permission component* is similar, but without constraints: the inputs are conditional permissions (coming from the norms base) and brute and institutional facts (coming from the counts-as component). The outputs are permissions. Thus, the whole architecture transforms brute facts into obligations and permissions.

Agent cognitive capacities

In analysing this architecture it is striking that first, the input of the whole architecture is restricted to brute facts and secondly, that the norms base has no input.[10] At first sight this seems to be straightforward. However, this leaves a problem unresolved: the norms base can only be updated off-line. Norms are thus *not* emergent features of the system; there exist no feedback loop between the environment and the norms-base.

It can be presumed that this is due to the fact that sensory data consist of brute facts. Based on ethnological evidence, Emil Durkheim [64] proposed an alternative view in his 'elementary form of religious life'. First, he claimed that religion has been the very first representation of a norm setting authority in the

[10] This is not specific to this example. In fact, also for the BDI inspired architectures it holds that belief updating is realised by sensory data. Moreover, in attempts to resolve conflicts between the different components it is a commonplace that beliefs override the other components. This is the architecture of so-called realistic agents [54]. In fact, the update of the obligation component is an open problem for the design of normative agents.

evolution of human culture. Religion entails belief formation affecting the norms base. For instance, if you convert to a muslim you are prohibited to drink alcohol. This is uncontroversial. However, Durkheim also claimed that religion always entails a cosmology. Religious belief is not only about supernatural entities but about the structure of the world. Hence, brute and institutional facts cannot be separated. An obvious objection is that here two meanings of belief are confused. It is a difference to belief that it is raining just now or to belief in the existence of a supernatural god. However, this is exactly what is contested by Durkheim's analysis: "Religious representations are collective representations which express collective realities" [64, p. 22].

Moreover, it is worth noting that they are *collective* representations: by analysing totemism Durkheim concluded that primitive religion had been the very first representation of society by its members. Totemism established the identity of the clan by the shared name of the totem and thus a relation between the individual and the society. The symbolic unity, generated by this classification scheme, was the source of this belief system to exert moral authority in much the same way as the belief that it rains might trigger the intention to open an umbrella.

Obviously, this is difficult to implement computationally. However, as the notion of a 'shared name of the totem' indicates, the problem is closely related to the emergence of *language* [comp. 65]. The advent of language opens up the possibility of a symbolic representation of the world in a consensual linguistic domain [35]. Language is thus a precondition for religious collective representations. However, *this* is a computationally traceable problem. For instance, the emergence of a commonly shared lexicon is an implemented feature of agent models [66, 67]. Eventually, the current New-Ties Project might provide new insights "which system components carry the knowledge structures that make up world models".[11] Anyway, to represent the *human* social dynamic in agent systems it is necessary to close the feedback loop between generating and emergent phenomena. Since the only environmental input in the agent's design is directed to the belief component (here the counts-as conditionals), this problem is in some way related to the relation between the update of beliefs and obligations. In the following we will see that beliefs can exert social control (similar to obligations).

5 SOCIAL CONTROL

While the former section was concerned with the question of how social norms work *in* the minds of individual actors, this section is concerned with the question of how they get *into* the minds. This is closely related to the problem of social control. In particular the invention of 'Artificial Virtues' [52] in large and anonymous societies demands for social control.

Sociologically, the problem of social control refers to the problem discussed above: the emergence of social positions. It has already been remarked that, presumably, this process was driven by population concentration. Hierarchical organisation of society allows for larger populations than prehistoric bands. Evolutionary psychology estimates that ancestral hominids lived in groups of 20 to 100 persons [68]. In small groups social control can be exerted in direct peer-to-peer interaction. In larger societies, however, this becomes precarious because actors can preserve anonymity.[12] A possibility to assert social order are new organisational features such as hierarchies and role differentiation: namely, by vesting hierarchical positions with norm setting authority. As a matter of fact, this process occurred in course of cultural evolution. In large societies, social control is executed by specialised institutions, providing specific social positions. [13]

Agent models of social control

In the following, an architecture of such a normative control system is investigated, described in the paper 'Norm governed multiagent systems: the delegation of control to autonomous agents' [69]. It has been stressed by several authors that normative obligations involve both – a norm addressee and someone wanting the norm to be fulfilled [70 – 73]. However, this architecture is selected for closer inspection, since it explicitly introduces agents with specialised social *roles*.

In the paper it is distinguished between 1) agents whose behaviour is governed by norms, 2) so-called defender agents that monitor violations and 3) a normative system that issues norms and monitors the defender agents. This reflects that in modern states the government is separated into a legislative system, responsible for the norm setting and a judicial system, responsible for the control of norm compliance. Thus, there exist three classes of autonomous agents: (class of) agent 1 is subject of obligations, (class of) agent 2 is responsible for norm control and sanctioning of violations. Agent 2 is the defender agent. (class of) agent 3 is the central authority that imposes obligations and permissions and monitors the defender agents. All agents make their decisions autonomously, i.e. based on their interests and states of belief.

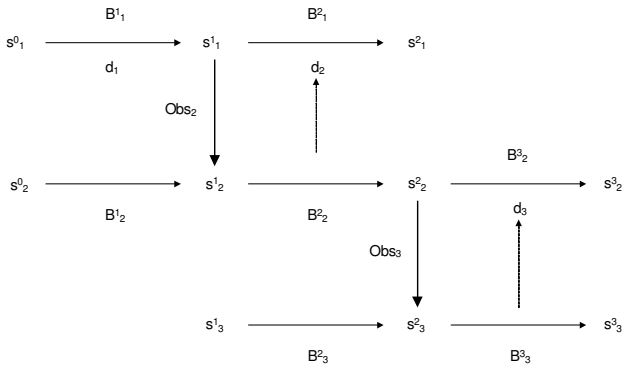

Figure 4. (Three agent) Control System

The sequence of actions in a three agent scenario is the following (from the perspective of agent 1):

Agent 1 makes its decision d_1 at time 0. It beliefs to be in a state s^0_1 (subscripts denote the agents, superscripts denote time).

[11] **http://www.cs.vu.nl/~gusz/newties/newties.html**

[12] This refers to the well known problem of cooperation among strangers. It shall not be questioned that this problem also can be resolved by decentralised interaction [74].

[13] The problem that the generation of a certain surplus is a presupposition to provide specialised positions will not be investigated in this paper [comp. 75].

The expected consequences of the decision at time 1 are calculated by a belief rule B^1_1. The expected consequences are denoted as the epistemic state s^1_1 of the agent 1. In making its decision, agent 1 tries to take the decision of agent 2 into account. Therefore it has a representation of what it beliefs to be agent 2's initial state s^0_2. Agent 1 beliefs that its decision has the consequence that agent 2 then beliefs to be in the state s^1_2, since agent 2 observes agent 1. Then agent 2 makes its decision d_2. The decision is based on whether agent 2 counts the action of agent 1 as a norm violation or not. Thus, (next to goals and desires) agent 1 builds its decision on expectations about the belief system of agent 2.

However, the same holds for agent 2: At time 1 (when agent 2 observes agent 1), agent 2 beliefs agent 3 to be in the epistemic state s^1_3. Moreover, it beliefs that the epistemic state of agent 3 changes to s^2_3 as a consequence of its decision d_2. This in turn will cause a decision of agent 3. Agent 2 beliefs that this will lead to the epistemic state s^3_2 for itself and to s^3_3 for agent 3. Thus, (next to goals and desires) agent 2 builds its decision on expectations about the belief system of agent 3.[14]

The decision process is characterised as follows: the agents are assumed to be of a selfish stable agent type. That is, it is not implemented that agents automatically obey norms, but calculate an optimal decision. An optimal decision maximises expected utility.

Obligations, i.e. norms, are characterised as follows: Agent i beliefs that it is obliged to do x with sanction s under condition q if it beliefs that agent i+1 (the defender agent) desires and has the goal x. Moreover, agent i beliefs that agent i+1 has the desire not to sanction and agent i itself has the desire not to be sanctioned. However, agent i beliefs that agent i+1 has the desire that there is no norm violation and has the goal and desire to recognise a norm violation. Finally, agent i beliefs that agent i+1 desires to sanction a norm violation if it recognises it.

Note, that defender agents do not intrinsically desire to sanction. Defender agents desire to sanction a norm violation because they are monitored by the norm setting authority. The norm setting authority is represented by agent i+2. This means (from the perspective of agent i) that agent i beliefs that it is obliged to do x, if it beliefs that agent i+2 desires and has the goal that x and that there is no norm violation. Moreover, agent i beliefs that (agent i+2 beliefs that) agent i+1 is conditionally obliged by agent i+2 to sanction a violation by agent i. The obligation for the defender agent to sanction norm violation is again represented in the same way: namely by the possibility that the norm setting authority, agent i+2, sanctions violations of the obligation to sanction norm violation. Hence, defender agents sanction because of fear of sanctions.[15]

[14] In principle, this can be iterated: for example, the central authority (agent 3) can delegate the control of the defender agent (agent 2) to another defender agent. This leads to a hierarchy of agents in which each agents considers the reaction of the agent in the subsequent hierarchy level when choosing an action. Conversely each agent observes the agents on the next lower hierarchy level.

[15] In principle, this is the same structure than in Robert Axelrod's so-called meta-norms game [76]. In this game agents can sanction defecting agents and agents not sanctioning defections. The difference is, that role differentiation is introduced here. Sanctioning norm violations and monitoring sanctions is ascribed to specific types of agents.

Orders of emergence

In this architecture norm enforcement is ascribed to specific types of agents. It thus explicitly represents social *role differentiation*. This implies the existence of social positions assigned to specific tasks. It is worth noting, that first, defender agents are obliged to sanction because of their professional duties and secondly, that this role differentiation is a feature of the agents' *belief* structure. Social positions are represented as a feature of the agents' cognitive capacities.

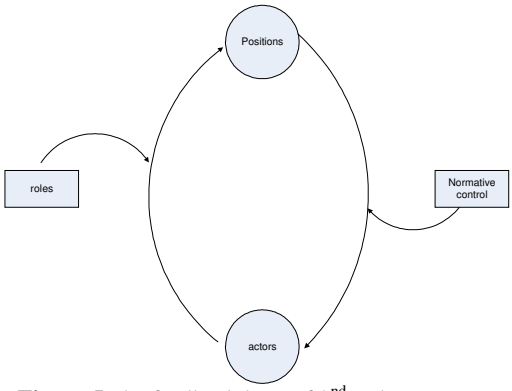

Figure 5. the feedback loop of 2nd order emergence

The stabilisation of the process of norm *immergence* by means of social control (in large societies) can be enforced by the *emergence* of social positions, enabling the delegation of control. The notion of positions refers to the emergence of hierarchical organisation of society as analysed in the EOS model [45]. Hence, both processes of emergence and immergence refer to each other. The delegation of social control to specific agent roles closes the feedback loop between emergent and immergent processes. It enables the possibility of the invention of 'Artificial Virtues'. However, currently the feedback loop is not closed: the existence of positions is a pre-given feature of this architecture. Only an integrated perspective on all three cases enables a representation of a complex feedback loop of social emergence.

6 CONCLUSION & FUTURE WORK

The task of this paper was to argue that the cognitive structure of software agents opens-up a new perspective to an explanation of social emergence. Agent-based models and architectures have been investigated with regard to the question of how individual agents are able to produce and to be in the same time a product of social reality. A three-stage scenario of the evolution of social reality has been developed. Particular attention has been paid to the question how society gets effective by the agents' cognitive structure. This enables to develop a theoretical framework that is capable to represent simultaneously the social micro- and macro-level.

The process of differentiation of social positions and individual actors has been identified as the *emergence* of social reality. The EOS model [45] demonstrates how social structure can be represented in the agents' social model. However, it still lacks a process of abstraction of social positions from individual

agents. The representation of counts-as conditionals in the architecture developed by Boella/van der Torre [62] opens-up a perspective of how emergent social reality can be *causally effective* in the individual agents' minds. By the generation of permissions and obligations, society is reproduced again. Also here, society is a component of the agents' cognitive structure. However, the shortcoming of this architecture is that the norms base can only be updated off-line. If agents would be able to develop conjointly a symbolic representation of the world in a consensual linguistic domain, it could be possible that a norms base could be developed in this process. The processes of emergence of social positions and immergence of social norms are related by norm enforcement exerted by actors ascribed to *specific roles* (that is, social positions), *transferring* social norms into individuals. The principles of role differentiation are described in Boella/van der Torre's paper [69]. Note, that these roles are features of the agents' belief structure. In this model, however, positions (of defender agents and normative authority) are pre-given.

So far the models stand in isolation. It is an open problem to link the models recursively. The task would be to integrate a model of the emergence of positions, responsible for the execution of social control, with a model of norm internalisation: if agents would be able to recognise the emergent norm setting authority, a complex feedback-loop of 2nd order emergence would be established. It is thus left for future work to close the feedback-loop and to develop a more comprehensive model of emergence in the loop.

ACKNOWLEDGEMENT

This work has been undertaken as part of the Project 'Emergence in the Loop' (EmiL: IST-033841) funded by the Future and Emerging Technologies programme of the European Commission, in the framework of the initiative "Simulating Emergent Properties in Complex Systems".

REFERENCES

[1] K. Knorr-Cetina, and A. V. Cicourel (Eds.), *Advances in social theory and methodology*. London: Routledge (1981).

[2] M. Archer, *Realist social theory: the morphogenetic approach.* Cambridge: Cambridge University Press (1995).

[3] R. Mayntz, Individuelles Handeln und gesellschaftliche Ereignisse: Zur Mikro-Makro-Problematik in den Sozialwissenschaften. MPIfG Working Paper, 99/5 (1999).

[4] K. Sawyer, Artificial Societies: Multiagent Systems and the Micro-Macro Link in Sociological Theory. *Sociological Methods and Research,* 31/3 (2003).

[5] B. Heintz, Emergenz und Reduktion. Neue Perspektiven auf das Mikro-Makro-Problem. *Kölner Zeitschrift für Soziologie und Sozialpsychologie,* 56/1 (2004).

[6] M. Brodbeck (Ed.), *Readings in the Philosophy of the Social Sciences.* New York: MacMillan (1971).

[7] J. O'Neill (Ed.), *Modes of Individualism and Collectivism.* London: Heinemann (1973).

[8] R. Bhaskar, *The Possibility of naturalism: A philosophical critique of the contemporary human sciences.* (Second Edition) Hemel Hempstead: Harvester (1989).

[9] R. Collins, On the Microfoundations of Macrosociology. *American Journal of Sociology,* 86 (1981).

[10] A. Giddens, *The constitution of Society: Outline of the Theory of Structuration.* Cambridge: Polity Press (1984).

[11] K. Sawyer, *Social Emergence: Societies as Complex Systems.* Cambridge: Cambridge University Press (2005).

[12] M. Neumann, Emergence as an Explanatory Principle in Artificial Societies, in: *Proceedings of the Conference on Epistemological Perspectives on Simulation II,* K. G. Troitzsch, F. Squazzoni (Eds.). Sythese Library, Berlin: Springer (forthcoming).

[13] A. Drogul, J. Ferber, Multi-Agentsimulation as a tool for studying emergent processes in societies. In: *Simulating Societies. The computer simulation of social phenomena.* N. Gilbert, J. Doran (Eds.). London, UCL Press (1994).

[14] G. Deffuant, S. Moss, W. Jager, Dialogues Concerning a (possibly) new Science. *Journal of Artificial Societies and Social Simulation* 9/1 (2006). `http://jasss.soc.surrey.ac.uk/9/1/1.html`

[15] A. Beckermann, H. Flor, J. Kim, (Eds.) *Emergence or Reduction?* Berlin, New York: De Gruyter (1992).

[16] J. Holland, *Emergence - From Chaos to Order.* Oxford, Oxford University Press (1998).

[17] J. A. Goldstein, Emergence as a construct: History and Issues, *Emergence* 1 (1999).

[18] V. Darley, Emergent Phenomena and Complexity, In: R. Brooks, P. Maes (Eds.). *Artificial Life IV: Proceedings of the Fourth International Workshop on the Sythesis and Simulation of Living Systems.* Cambridge, Mass.: MIT Press. (1994).

[19] K. Richardson, On the Limits of bottom-up computer simulation: Towards a nonlinear modelling culture, in: *Proceedings of the 36th Hawaiian International Conference on System Science,* IEEE, California (2003).

[20] C. Emmeche et al., Explaining Emergence: Towards an Ontology of Levels, *Journal for General Philosophy of Science* 28 (1997).

[21] E. Mayr, *Eine neue Philosophie der Biologie,* München, Piper (1991).

[22] E. Thomson, F. Varela, Radical Embodyment: neural dynamics and consciousness, *Trends in Cognitive Science* 5 (2001).

[23] W. Sullis, Archetypical dynamical systems and semantic frames in vertical and horizontal emergence, *Emergence: Complexity and Organisation,* 6/3 (2004).

[24] M. Pauen, G. Roth (Eds.), *Neurowissenschaften und Philosophie.* Wilhelm Fink Verlag, München (2001).

[25] R. Oerter Entwicklungskrisen im Jugendalter: Eine Systemtheoretische Perspektive, *Psychotherapie Heft 1* (2002).

[26] Y. Engeström, R. Mietinen, R.L. Punamäki (Eds.) *Perspectives on activity theory.* Cambridge: Cambridge University Press (1999).

[27] M. Bedau, Weak Emergence. In: *Philosophical Perspectives: Mind, Causation, and World*, Vol. 11. J. Tomberlin (Ed.). Malden MA: Blackwell (1997).

[28] N. Gilbert, Varieties of emergence. Paper presented at the social Agents: Ecology, Exchange and Evolution Conference on Social Agents: ecology, exchange and evolution. Chicago (2002).

[29] C. Castelfranchi, Through the Minds of the Agents. *Journal of Artificial Societies and Social Simulation,* 1/1 (1998). `http://www.soc.surrey.ac.uk/JASSS/1/1/5.html`

[30] G. Andrighetto, M. Campennì, R. Conte, M. Paolucci, On the Immergence of Norms: a Normative Agent Architecture. In: *Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence,* Washington DC (2007).

[31] D. Campbell, Downward Causation in Hierarchially Organised Biological Systems. In: *Studies in the Philosophy of Biology,* F. Ayala, T. Dobzhansky, (Eds.). London: MacMillan (1974).

[32] R. Conte, G. Andrighetto, M. Campennì, M. Paolucci, Emergent and Immergent Effects in Complex Social Systems. In: *Proceedings of the AAAI ,07,* (2007).

[33] J.S. Sichman, R. Conte, C. Castelfranchi, Y. Demazeau, A Social Reasoning Mechanism Based on Dependence Networks. In: *Proceedings of the 11th European Conference on Artificial Intelligence*, A.G. Cohn (Ed.) Baffin Lane, England: John Wiley and Sons (1994).

[34] D. Dennet, *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon and Schuster; London: Allen Lane (1995).

[35] C. Goldspink, R. Kay, Emergence in Social Systems: Distinguishing Reflexive and Non-reflexive modes. In: *AAAI Fall Symposium: Emergent Agents and Socialities: Social and Organizational Aspects of Intelligence*. Washington (2007).
[36] P. Blau, A Macrosociological Theory of Social Structure. *American Journal of Sociology,* 83/1 (1977).
[37] N. Luhmann, *Soziale Systeme: Grundriss einer allgemeinen Theorie,* Frankfurt a.M.: Surkamp (1984).
[38] P. Mellars, The ecological basis of social complexity in the Upper Palaeolithic of southwestern France. In: *Prehistoric hunter-gatherers: the emergence of cultural complexity.* T. Price, J. Brown (Eds.). New York: Academic Press, (1985).
[39] J. Aigner 1989, Frühe Siedlungen im arktischen Nordamerika. In: *Siedlungen der Steinzeit,* Heidelberg: Spektrum, (1989).
[40] M. Kolb, Monumental Grandeur and the rise and fall of Religious Authority in Precontact Hawaii. *Current Anthropology,* 34 (1994).
[41] C. Renfrew, Symbol before Concept: Material Engagement and the Early Development of Society. In: *Arachaeological Theory Today,* I. Hodder (Ed.). Cambridge, Polity Press, (2001).
[42] T. Earle, Culture Matters in the Neolithic Transition and Emergence of Hierarchy in Thy, Denmark: Distinguished Lecture. *American Anthropologist,* 106, (2004).
[43] C. Kresl, E. Van West, Carr, R. Wilshusen, Be There Then: A Modeling Approach to Settlement Determinant and Spatial Efficiency among late Ancestral Pueblo Populations of the Mesa Verde Region, U.S. Southwest. In: *Dynamics in Human and Primate Societies: Agent-Based Modeling of Social and Spatial Processes.* G. J. Gumerman, T. A. Kohler (Eds.). Oxford: Santa Fe Institute and Oxford University Press (2000).
[44] A. Gumerman, J. Swedlund, S. Harburger, R. Chakravarty, J. Hammond, J. Parker, M. Parker, Population Growth and Collapse in a Multi-Agent Model of the Kayenta Anasazi in Long House Valley. In: *Proceedings of the National Academy of Sciences of the United States of America* 99, 3 (2002).
[45] J. Doran, M. Palmer, N. Gilbert, P. Mellars, The EOS Project: modelling Upper Palaeolithic social change. In: *Simulating Societies.* N. Gilbert, J. Doran (Eds.). London, UCL Press, (1994).
[46] M. Cohen, Prehistoric Hunter-Gatherers: The Meaning of Social Complexity. In: *Prehistoric hunter-gatherers: the emergence of cultural complexity,* T. Price, J. Brown (Ed.). New York: Academic Press, (1985).
[47] N. Gilbert, Emergence in social simulation. In: *Artificial Societies: The computer simulation of social life,* N. Gilbert, R. Conte (Eds.). London: UCL Press, (1995).
[48] G.H. Mead, *Mind, Self, and Society.* Edited by C.W. Morris. Chicago: University Press (1934).
[49] T. Parsons, *The Structure of Social Action. A Study in Social Theory with Special Reference to a Group of Recent European Writers.* New York, London: Free Press, (1968 [1937]).
[50] T. Parsons, E.A. Shils, *Towards a General Theory of Action.* Harvard: Harvard University Press, (1951).
[51] R. Darendorf, *Homo Sociologicus. Ein Versuch zu Geschichte, Bedeutung und Kritik der Kategorie der sozialen Rolle.* Opladen: Westdeutscher Verlag (1956).
[52] D. Hume, *A Treatise in Human Nature,* Vol 2, Edinburg Edition, (1826).
[53] A. Rao, M. Georgeff, Modelling rational agents within a BDI architecture. In: *Proceedings of the KR 91,* (1991).
[54] J. Broersen, M. Dastani, J. Hulstijn, Z. Huang, L. van der Torre, The BOID Architecture: Conflicts between Beliefs, Obligations, Intentions, and Desires. In: *Proceedings of the 5th International Conference on autonomous agents,* (2001).
[55] J. Broersen, M. Dastani, L. van der Torre, Beliefs, Obligations, Intentions, and Desires as Components in an Agent Architecture. *International Journal of Intelligent Systems,* 20 (2005).
[56] G. Boella, Deliberate normative Agents. Basic Instructions. in: *Social Order in Multiagent Systems.* R. Conte, C. Dellarocas (Eds.). Norwell: Kluwer (2001).
[57] A. Garcia-Cumino, A. Rodriguez-Aguilar, C. Sierra, W. Vasconcelas, Norm-oriented programming of electronic institutions. *AAMAS '06,* (2006).
[58] F. Dignum, D. Kinny, L. Sonenberg, From desires, obligations and norms to goals. *Cognitive Science Quarterly,* Vol. 2 (2002).
[59] F. Lopez y Lopez, M. Luck, M. d'Inverno, Constraining autonomy through norms. *AAMAS '02,* (2002).
[60] J. Vazquez-Salceda, H. Aldewereld, F. Dignum, Norms in Multi-Agent Systems: From theory to practice. *International Journal of Computer Systems and Engineering,* 20 (2005).
[61] C. Castelfranchi, F. Dignum, C. Jonker, J. Treur, Deliberative normative agents: Principles and Architecture. In: *Intelligent Agents: Theories, Architectures, Languages.* Springer, Berlin (2000).
[62] G. Boella, L. van der Torre, An Architecture of a Normative System. *AAMAS '06*, ACM Press (2003).
[63] J. Searle, *The construction of social reality.* London: Penguin Press, (1995).
[64] E. Durkheim, *The Elementary Forms of Religious Life.* New York: Free Press (1965[1912]).
[65] H. Whitehouse, Modes of religiosity: towards a cognitive explanation of the sociopolitical dynamics of religion. *Method and Theory in the study of religion,* 14 (2002).
[66] L. Steels, Constructing and Sharing Perceptual Distinctions. Paper presented at the European Conference on Machine Learning: Berlin, (1997).
[67] T. Gong, J. Ke, J. Minett, W. Wang, A computational Framework to simulate the Co-evolution of language and Social Structure. In: *Artificial Life IX.* Boston, (2004).
[68] J. Tooby, L. Cosmides, Conceptual Foundations of Evolutionary Psychology. In: *Handbook of evolutionary psychology.* D. Buss (Ed.). Hoboken: Wiley, (2005).
[69] G. Boella, L. van der Torre, Norm Governed multiagent systems: the delegation of control to autonomous agents. In: *Proceedings of the IEEE/WIC IAT Conference,* IEEE Press (2003).
[70] R. Conte, C. Castelfranchi, From Conventions to prescriptions: Towards an integrated view of norms. *Artificial Intelligence and Law,* 7 (1999).
[71] G. Boella, L. Lesmo, Deliberate Normative Agents. In: *Social Order in Multiagent Systems.* R. Conte, C. Dellarocas (Eds.) Norwell: Kluwer, (2001).
[72] R. Conte, F. Dignum, From Social Monitoring to Normative Influence. *Journal for Artificial Societies and Social Simulation,* 4/2 (2001). `http://www.soc.surrey.ac.uk/JASSS/4/2/7.html`
[73] F. Lopez y Lopez, A. A. Marquez, An Architecture for Autonomous Normative Agents. In: *Proceedings of the Fifth Mexican International Conference in Computer Science ENC '04,* IEEE (2004)
[74] R. Schüssler, *Kooperation unter Egoisten: 4 Dilemmata.* München: Oldenbourg, (1990).
[75] N. Müller, *Civilization Dynamics, Vol. 1: Fundamentals of a model oriented description.* Aldershot: Avebury, (1989).
[76] R. Axelrod, An evolutionary approach to norms. *American Political Science Review,* 80 (1986).

What can Agents Know? The Feasibility of Advanced Cognition in Social and Economic Systems

Paul Ormerod[1]

Abstract The purpose of this paper is to suggest that in many social and economic contexts, self-awareness of agents is of little consequence. The complexity of many such systems is very high. No matter how advanced the cognitive abilities of agents in abstract intellectual terms, it is as if they operate with relatively low cognitive ability within the system. This can be the case even when the emergent properties of the system are known to individual agents. Examples are given from macro-economics, the evolution of firms, financial markets and games.

1 INTRODUCTION

The purpose of this short paper is to suggest that in many social and economic contexts, self-awareness of agents is of little consequence. The complexity of many such systems is very high. No matter how advanced the cognitive abilities of agents in abstract intellectual terms, it is as if they operate with relatively low cognitive ability within the system. This can be the case even when the emergent properties of the system are known to individual agents.

This is not to say that the arguments of Gilbert [1] and of Goldspink and Kay [2] are not valid in some empirical settings. But I want to suggest that the situations in which they are valid might be rather circumscribed. The Turing rule that the vast majority of real life problems have no algorithmic solution limits the empirical usefulness of the assumption that agents operate with advanced cognitive ability. In many real life situations, the dimension of the problem scales super-exponentially, even when considering situations in which interactions between agents and emergent properties of the system are absent. Keen [3], for example, provides an illustration of the dimensions involved in even quite simple consumer choice decisions.

I am not pretending to offer in any way a proof of my main point. But I want to provide empirical examples. I am suggesting that a key issue in the context of the theme of this session of the conference is prior consideration of the feasibility of agents exhibiting advanced cognition in any particular context, regardless of their inherent intellectual abilities.

2 THE MACRO-ECONOMY

The problems of assigning advanced cognition to agents, even with no emergence, can be illustrated first of all by reference to the macro-economy.

The economy is undoubtedly a complex system with emergent properties. The decisions of millions of individual consumers and firms interact to produce the movements we observe at the aggregate, macro-economic level in variables such as total output (GDP), inflation and unemployment.

Policy makers have a strong incentive to be in possession of forecasts which are systematically accurate over time. If they have little idea of where the economy is likely to be in a year's time, say, the ability to carry out a successful policy intervention is obviously limited.

Policy makers and their advisers are very much aware of the emergent phenomena of a macro-economy. Indeed, it is precisely this data, which emerges from the interactions of the agents at the micro level, which they are trying to influence.

There is a large literature which shows that even on one-year ahead predictions, the forecast errors are large relative to the size of the data. Forecasters seem to do well when the economy is pretty stable, but are quite unable to capture turning points. The evidence from the forecasting record suggests that we can do ever so slightly better than the naïve rule which says next year's growth is the same as this, but there is not much in it. The only economy which seems to deviate from this finding to any significant extent is that of the United States, yet even here the level of predictability is very low by scientific standards.

There is a whole branch of economic theory, real business cycle theory, which argues precisely that cycles arise from random exogenous shocks. As it happens, I think the cycle is mainly endogenous and not exogenous. But here is a serious part of mainstream theory which hypothesises that the short-term growth rate of the economy is unpredictable. Of course, predictions can always be carried out, but 'unpredictable' here means that it is not possible to make systematically accurate forecasts. So our ability to learn to make short-term macro forecasts is very severely constrained. We do not appear to be able to process successfully the available information.

I started off many years ago as a forecaster, and it soon became clear to me that it didn't work. I've been interested for a long time in why this should be the case, and a few years ago I finally worked it out. Physicists and mathematicians have developed a technique, random matrix theory, which enables us to decompose time series data into what we might usefully think of as signal and noise. Signal is the bit that contains genuine information, and noise is, well, noise. I used this technique and published an article in *Physica A* on the failure of macro-economic forecasting [4]. Essentially, the data is dominated by noise rather than signal.

The lack of predictability of the cycle does not mean that the agents taking the decisions which generate the cycle are acting at random. It is a question of dimensionality. The dimension of the problem leads to the data appearing 'as if', a favourite phrase of economists, it is close to random.

[1] Volterra Consulting, London and Institute of Advanced Study, University of Durham Email: pormerod@volterra.co.uk

Theoretical approaches which contain sophisticated autonomous agents following maximising rules, such as real business cycle theory, are unable to capture key emergent features of the system, such as the (weak) time and frequency domain regularities, the distribution of the duration and size of recessions, the overall distribution of GDP growth rates and the distribution of individual firm growth rates. In contrast,. models which have simple agent rules, but in which agents are connected on networks, can.[for example, 5]

A series which, at least for the UK, is not possible to distinguish from a random one is the *change* in the rate of inflation [6] The Monetary Policy Committee of the Bank of England have as their objective the maintenance of a particular rate of inflation. All the members are thoughtful and intelligent. They have large teams of highly qualified researchers employed to discover the emergent properties of the system, in other words how the macro-economy operates.

Yet, given that the change in inflation is indistinguishable from a random series, they do not know what the rate of inflation is going to be in, say, one year's time. Specifically, they do not know whether it will be higher or lower than it is at present. So their ability to control the rate of inflation, to meet the target, is very seriously constrained[2].

The macro-economy is an important example of our inability to learn in any meaningful sense because of limits to the cognitive ability of agents. It is not that the agents – policy makers – do not have advanced cognition in general, it is as 'as if' in specific contexts they do not.

3 FIRM EVOLUTION AND DEATH

Firms have a very strong incentive to survive. Management spends a large amount of both time and money in trying to understand the properties of the system which their particular firm inhabits. Yet even amongst the very largest firms in the world, success tends not to persists. Batty [7] notes that over the 1955-1994 period 'From the 100 firms making up the Fortune 500list in 1955, 39 (percent) remain in 1994, and if the changes in each year from 1955 are examined, this reveals a more dramatic micro-dynamics with firms entering and leaving the list with great rapidity'.

Even more dramatically, firms both large and small actually disappear all the time. In both America and Europe, for example, more than 10 per cent of firms become extinct every year. A stylised fact that is established in the literature relates to the relationship between the age of the firm and the probability of extinction, or survival, looking at it from the opposite perspective. The probability of extinction is high in the early years of a firm's life. It falls rapidly, and becomes essentially flat. The finding seems to apply to firms regardless of size.

The second relates to the concept of the size of an extinction event. In other words, during a given period of time, we identify what proportion of firms become extinct. This gives us the size of the event. The bigger the proportion, the bigger the size. We then relate the size to the frequency with which it is observed over the whole of the data.

[2] Even assuming economists understand the connection between changes in interest rates now (the policy instrument available to the Bank) and changes in the rate of inflation in the future, which they do not. For example, the impact of interest rate changes on the exchange rate, which is a determinant of inflation, is theoretically indeterminate

This approach is well known in the biological fossil record. The frequency with which any given size of extinction event is observed – in this case the proportion of all species becoming extinct during given time interval – is inversely proportional to the square of the size. In other words, large extinction events are much less frequent than small ones.

It turns out that a very similar relationship exists for the frequency-size relationship observed for the extinction of firms [8]. So we have two stylised facts to explain.

I developed a theoretical model to account for them [9]. The actions of any given firm can have either positive or negative net impacts on the fitness for survival of any other firm. In other words, firms might produce complimentary products, or they might compete. The impacts are expressed through a matrix, each cell of which specifies the impact of firm i on the fitness for survival of firm j. There is a rule which specifies what level of fitness a firm needs to survive, and how extinct firms are replaced.

The key feature of the model is that, in each step of the model, the net impact of the actions of firm i on the fitness for survival of firm j is updated at random. In other words, it is as if firms had no knowledge of the impact of their strategies. It is as if they are unable to learn from their past experience.

This model replicates very accurately the two key stylised facts on firm extinction. I go on to allow firms to have a certain amount of knowledge of the consequences of their actions. In other words, it is as if they are able to learn. I vary both the proportion of all firms in the model with this ability, and the amount which they are able to learn.

There are very considerable gains to being able to learn. The average life of agents able to learn, even small amounts, is considerably larger than agents who cannot learn. In the limit, of course, as the proportion of firms able to learn approaches unity, and as the amount of knowledge which they learn increases, the firms approach infinite lives and never die.

But firms are only able to know very small amounts before the model ceases to replicate the stylised facts on extinction. It appears that a firm cannot learn very much at all either about the impact of their strategies on the ability of other firms to survive, or about the impact of the strategies of other firms on its own ability to survive.

As with business cycle theory, the key to the ability of the theory to account for emergent behaviour is not the sophisticated rules of the agents, but the fact that they are connected on a network. This structure of the system is more important than specific agent rules. Knowledge of the emergent properties of the system would essentially be of no use to an individual firm.

4 FINANCIAL MARKETS

A similar result on structure is obtained by Farmer et. al. in their study of share prices on the London Stock Exchange [10] Agents place orders to buy and sell at random, subject to constraints imposed by current prices (which limit the size of order which can be placed by an individual). The model explains 96 per cent of the variance of the gap between the best buying and selling prices (the spread) using a sample of 11 stocks, and 76 per cent of the variance of the price diffusion rate, which determines the size and frequency of changes to prices.

The price setting mechanism is a continuous double auction, the process actually used on the Stock Exchange. The auction is

called "double," because traders can submit orders to both buy and sell, and "continuous," because they can do so at any time. The results of the Farmer et. al. paper arise, for reasons which no-one yet understands, because of this particular price setting mechanism which is used.

In real life, of course, agents engage in what they believe to be clever strategic behaviour, yet a model which neglects this entirely performs impressively from a scientific perspective. So, again, the *structure* of what we might think of as the game, the price setting mechanism, is very important in determining outcomes, more so than accurate modelling of agent behaviour.

Further, knowledge of the emergent properties of the system, the subtle properties of price changes, is of no use to individual agents in the system. They cannot exploit knowledge of these properties in their individual decision rules.

5 GAMES

The more abstract world of game theory offers further illustrations both of the difficulties of assigning high cognition to agents, and of the importance of the overall structure of the system, the institutional rules. But, again, knowledge of the 'emergent' properties of the game would not be of much help to individuals.

A trivially easy game is noughts and crosses (British English) or tic-tac-toe (American English). There are multiple Nash equilibria in the game, so many that almost any move in any strategy leads to the optimal outcome, a draw. Knowledge of this emergent property might be useful, but it is easy to discover by experiment. Even young children rapidly learn this. But we do not need to consider games which are very much more complicated before things become less clear cut.

The rules of chess can be stated very readily, and a reasonably intelligent person can remember them quickly. But the computational power required to analyse most position in a game scales super-exponentially. In the vast majority of positions which can exist, we are completely unable to determine which is the best move.

Computers have essentially made progress in chess by pure number crunching. In other words, by the exhaustive examination of permutations of moves in a given situation. The world's leading player for two decades after the Second World War, Botvinnik, believed that computers would eventually beat humans if they could in some way be programmed to understand the nuances of positional play in chess, rather than by exhaustive examination of the possibilities. He led a Soviet research programme on this, but essentially got nowhere. The gains have not been made by computers exhibiting advanced cognition in understanding the subtleties of positional play in chess – the emergent properties, as it were - but by grinding out tactical calculations.

So chess is an example of a game which can be described very simply, but where the dimension of the problem of solving it scales in a super-exponential way. Even very powerful modern computers can only solve a limited proportion of all possible 6 piece combinations yet the game itself involves 32 pieces.

Chess of course is a recreational game without wider applications. A game which is often held to have many practical application is that of the Prisoner's Dilemma. The rules are very simple and are time invariant. Agents are assumed to have a great deal of information. In particular, in its simplest form, each agent is assumed to know the payoff values of his or her opponent. This is a pretty strong assumption to make when you think about it. Yet do we know the best strategy? Well, we do when we make the very specific assumption that the game will end in a fixed number of moves, and that both players know this. We might think of this as removing any uncertainty which the existence of the future might bring. In other words, it limits the dimension of the problem.

There is a vast literature on the Prisoner's Dilemma when it ends at random. But the optimal strategy remains unknown. Agents do not have the cognitive ability to compute it. The scientific community has invested a great deal of effort in trying to discover, to learn the best strategy, but still we do not know.

The Beauty Contest game is based on Keynes' famous comment on the stock market, which he likened to a newspaper game popular in the UK in the 1930s. Newspapers published picture of 100 women, and to win it was necessary to guess the 6 which the most participants would select as the most beautiful. As Keynes wrote [11] 'It is not a case of choosing those [faces] which, to the best of one's judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest. We have reached the third degree where we devote our intelligences to anticipating what average opinion expects the average opinion to be. And there are some, I believe, who practice the fourth, fifth and higher degrees'

In the modern version, a group of individuals is asked to select a number between (and including) 0 and 100. The winner is the person whose guess is nearest to a specified fraction of the average of all the guesses. If all players practice a 'high degree' of reasoning, the winning number will be close to the Nash equilibrium of 0.

But experiments designed to elicit the degree of reasoning which agents use all show that it is low, typically between 1 and 3. For example, Duffy and Nagel [12] set up a game in which the winner is the person(s) whose guess was closest to half of either the median, the mean, or the maximum number chosen by all players. They found that players used a very low order degree of reasoning when forming expectations on other players' expectations. If the winning number were announced and the game repeated with the same players, they found that the winning number did approach zero, but even after repeated plays of the game the degree of reasoning remained low.

Knowledge of the Nash equilibrium solution would, except possibly in the final stages of successive plays of the game by the same set of players, be of no use to an individual. Indeed, anyone using this as his or her rule would in general lose. The key to success is not knowledge of the emergent equilibrium, but guessing the degree of reasoning which is being used by other players.

Interestingly, Duffy and Nagel found that the structure of the game, in this case the statistic which determines the winning number, had an important influence on the speed with which players converged towards zero as the winning number.

The Ultimatum Game [13] offers an example of how the wider setting in which a game is played can be of crucial importance to the outcome. Two players interact once only. The first player proposes how to divide a sum of money between themselves, and the second player can either accept or reject this proposal. If the second player rejects, neither player receives anything. If the second player accepts, the money is split according to the proposal.

There is now a vast literature on this game, almost rivalling that of the Prisoner's Dilemma. The reason for this is well known. Theoretically, the first player should offer the smallest non-zero amount possible, since from the point of view of the second player anything is better than nothing[3] . Yet the evidence that in general people do not follow this strategy is very strong. The game has been played in many settings, including ones in which the amount of money on offer is large to the participants, and this result appears to hold.

I am not suggesting that we know for certain why this is the case. But broader concepts of equity and fairness are surely important, 'broader' here in the sense of existing quite independently of the game itself. How agents play the game does not influence these broader sets of values, but the values influence the outcome of the game.

So, clearly, there are 'higher levels' of emergence which are important for outcomes. With the Beauty Contest game, it is the policy maker (the experimenter) who selects the statistic which is to be used. And with the Ultimatum Game, the outcomes appear to be determined by the broader values of society.

6 BRIEF COMMENTS

- In many social and economic contexts, self-awareness of agents is of little consequence. The complexity of many such systems is very high. No matter how advanced the cognitive abilities of agents in abstract intellectual terms, it is as if they operate with relatively low cognitive ability within the system. This can be the case even when the emergent properties of the system are known to individual agents.
- The Turing rule that the vast majority of real life problems have no algorithmic solution limits the empirical usefulness of the assumption that agents operate with advanced cognitive ability.
- The more useful 'null model' in social science agent modelling is one close to zero intelligence. It is only when this fails that more advanced cognition of agents should be considered.

REFERENCES

[1] N. Gilbert, *Varieties of Emergence.* Paper presented at the Social Agents: Ecology, Exchange, and Evolution Conference Chicago, (2002).

[2] C. Goldspink and R. Kay, *Social Emergence: Distinguishing Reflexive and Non-reflexive Modes* Paper presented at the AAAI Fall Symposium Washington (2007).

[3] S.Keen, *Debunking Economics: The Naked Emperor of the Social Sciences*, Sydney: Pluto Press (2001).

[4] P Ormerod and C Mounfield, 'Random Matrix Theory and the Failure of Macro-economic Forecasting,', *Physica A*, (2000).

[5] H Li and T Gao, 'A GDP fluctuation model based on interacting firms'. *Physica A*, forthcoming (2008).

[6] P Ormerod, 'Why Membership of the Monetary Policy Committee is a Sinecure', *Manchester Statistical Society*, (2001).

[7] M Batty, 'Visualising Creative Destruction', *Centre for Advanced Spatial Analysis, University College London, Working Paper 112*, (2007)

[8] C Di Guilmi, M Gallegati and P Ormerod, 'Scaling invariant distributions of firms' exit in OECD countries' *Physica A*, **334**, 267-273 (2004).

[9] P Ormerod and B Rosewell, 'What Can Firms Know?', *Proc. NACCSOS Conference*, Pittsburgh, (2003).

[10] J.D. Farmer, P. Patelli and I.I. Zovko, 'The Predictive Power of Zero Intelligence in Financial Markets', *Proc. Nat. Academy of Sciences*, **102**, 2254-59, (2005).

[11] JM Keynes, The General Theory of Employment, Interest and Money, *chapter 12, Macmillan*, (1936).

[12] J Duffy and R Nagel, 'On the Robustness of Behaviour in Experimental "Beauty Contest" Games', *Economic Journal*, **107**, 1684-1700, (1997).

[13] W. Guth, R. Schmittberger, and B. Schwarze , 'An Experimental Analysis of Ultimatum Bargaining', *Journal of Economic Behavior and Organization*, 3, 367-88, (1982).

[3] There is an argument as to whether offering zero itself to the second player also constitutes a Nash equilibrium

Agent Cognitive capabilities and Orders of Emergence: critical thresholds relevant to the simulation of social behaviours.

Chris Goldspink[1], Robert Kay[2]

Abstract In this paper we provide a brief recount of alternative approaches to what we argue is a fundamental issue for our understanding of sociality – the micro-macro problem or, as we refer to it here, the problem of social emergence. We then discuss recent attempts to identify how the range and type of emergent phenomena changes as a result of changes in the fundamental characteristics of micro-agents. We conclude that there appear to be a number of critical thresholds, notably that which arises when agents become constitutively autonomous and subsequently also develop behavioural (sensori-motor) autonomy. It is the combination of these two levels of autonomy which accounts for what we typically call 'cognition' in biological agents. Current artificial intelligence models attempt to replicate the ability without autonomy. While this approach is being seen as increasingly problematic in robotics it appears yet to have influenced approaches to social simulation. We propose achieving behavioural autonomy as a goal and focal point for future simulation research. We argue that this is the minimum threshold needed to achieve social emergence. We illustrate this by discussing the concept of social 'norm' as an 'attractor' in a phenomenal domain of structurally coupled behaviour..[12]

1 INTRODUCTION

Building and working with artificial societies using the methods of multi-agent social simulation serves us in several ways – it allows us to operationalize social theories and to compare simulated behaviours with those observed in the real world and it allows us to build new theory by exploring the minimal mechanisms that might explain observed social behaviour. Most importantly, it provides a unique ability to explore the interplay between levels of phenomena and to understand dynamic properties of systems. A great deal can and has been achieved in both these areas with even the simple methods we currently have available. However, Keith Sawyer [1] has recently reminded us that, to date, we have worked with agents of very limited cognitive capability and that this necessarily limits the range and type of behaviour which can be explored. This echoes a sentiment made a decade ago by Christiano Castelfranchi [2] that social simulation is not really *social* until it can provide an adequate account of the implication of the feedback between macro and micro which becomes possible with higher cognitive functioning of social agents.

This paper examines the relationship between agent capability and orders of emergence in order better to define the critical thresholds which limit our capacity to simulate certain classes of social phenomena.

In many respects, developments in our capacity to simulate artificial societies have led us to confront anew a long-standing issue within social theory. This problem is variously referred to as the micro-macro problem, the problem of structure and agency or social emergence. This problem has been a long term focus of our collaboration [see 3, 4]. Over the past decade we have worked towards a theory of sociality which can provide a coherent and consistent account of the interpenetration (circular causality) of micro and macro phenomena – i.e. which can provide a substantive account of fundamental social generative mechanisms. No such theory currently exists. This current paper is a continuation of that work but also has its origin in one author's involvement with the EU funded project titled Emergence in the Loop (EMIL). Through EMIL we aim to a) provide a theoretical account of the mechanisms of normative self-regulation in a number of computer mediated communities b) specify the minimum cognitive processes agents require to behave in normative ways c) develop a simulator which can replicate the range and type of normative behaviour identified by the empirical research so as to further deepen our understanding of how and under what conditions normative self-regulation is possible.

2 A BRIEF RECOUNT OF THE PROBLEM

The notion of emergence has a long history. Unfortunately the concept remains ill defined ambiguous and contentious, leading to the criticism that it stands as little more than a covering concept – used when no adequate account or explanation exists for some unexpected phenomena. The origin of the concept has been attributed to George Henry Lewes, in 1875 [5]. It subsequently found wide adoption within the philosophy of science but has been advanced within four streams: *philosophy*, particularly of science and mind; *systems theory*, in particular complex systems; *social science* where it has largely been referred to under the heading of the micro-macro link and/or the problem of structure and agency; and more recently in theoretical biology, cognitive theory and robotics. Interestingly there has been relatively little cross influence between these streams.

1 Centre for Research in Social Simulation, Department of Sociology, University of Surrey, Guildford, GU1 7XH, UK, c.goldspink@surrey.ac.uk

2 Executive Director and co-founder of Incept Labs, Sydney, Australia, rkay@inceptlabs.com.au

The Contribution from Philosophy of science

The philosophy of science and philosophy of mind stream is arguably the oldest – some date it back to Plato [6] but the debate is widely seen as having come to focus with the British Emergentists [7-9]. This school sought to deal with the apparent qualitatively distinct properties associated with different phenomena (physical, chemical, biological, mental) in the context of the debate between mechanism and vitalism. This stream remains focused on explaining different *properties* of classes of natural phenomena and with the relationship between brains and minds [See 10 for a recent summary of the positions]. As a consequence this has been the dominant stream within artificial intelligence. Peterson [6: 695] summarizes the widely agreed characteristics of emergent phenomena within this stream as follows. Emergent entities:

1. Are characterized by higher-order descriptions (i.e. form a *hierarchy*).
2. Obey higher order *laws*.
3. Are characterized by *unpredictable novelty*.
4. Are *composed of* lower level entities, but lower level entities are *insufficient* to fully account for emergent entities (*irreducibility*).
5. May be capable of *top-down causation*.
6. Are characterized by *multiple realization or wild disjunction* [11] (alternative micro-states may generate the same macro states).

Within this stream there is a concern with both upward and downward causation and it is the possibility for the later which attracts most argument. A key concept is *supervenience*: a specification of the 'loose' determinism held to apply between levels such that *'...an entity cannot change at a higher level without also changing at a lower level'* [12: 556]. Advocates of supervenience argue that properties associated by emergent structures exist only due to the properties of the underlying constituents and, in having no unique causal power other than those derived from those constituents, comprise only epiphenomena – they are not 'real'. This controversy persists within philosophical circles although it appears to derive in large part from an extreme form of physicalism [13]. Practicing physicists appear to have fewer problems with the concept than philosophers of mind. Physicists Clayton and Davies [10], for example, specify downward causation as involving macro structures placing *constraint* on lower level processes hence *'Emergent entities provide the context in which local, bottom up causation takes place and is made possible'* [6: 697]. Davies [14] argues that the mechanism of downward causation can usefully be considered in terms of boundaries. Novelty, he argues, may have its origin in a system being 'open'. He concludes:

> … top-down talk refers not to vitalistic augmentation of known forces, but rather to the system harnessing existing forces for its own ends. The problem is to understand how this harnessing happens, not at the level of individual intermolecular interactions, but overall – as a coherent project. It appears that once a system is sufficiently complex, then new top down rules of causation emerge (Davies 2006: 48).

For Davies then, top-down causation is associated with self-organization and may undergo qualitative transitions with increasing system complexity. For Davies also it is the 'openness' of some systems that 'provides room' for self-organizing process to arise, but he concludes, *'openness to the environment merely explains why there may be room for top-down causation; it tells us nothing about how that causation works.'* The devil then, is in the detail of the mechanisms specific to particular processes in particular contexts and particular phenomenal domains. Perhaps part of the problem with the concept is that it has been approached at too abstract a level.

The Contribution from Social Science

The micro-macro problem – the relationship between the actions of individuals and resulting social structures and the reciprocal constraint those structures place on individual agency – has long standing in social science. The problem is central to many social theories developed throughout the 19th and 20th century. Examples include: Marxian dialectical materialism [15] built upon by, among others, Vygotsky [16] and Lyont'ev [17]; the social constructionism of Berger and Luckmann [18]; Gidden's structuration theory [19]; and the recent work of critical realists [20-23]. These alternative theories are frequently founded on differing assumptions, extending from the essentially objectivist/rationalist theory of Coleman [24], through the critical theories of Habermas and then to the radical constructivism of Luhmann [25, 26].

Fuchs & Hofkirchner [27: 33] have recently suggested a four category schema for classifying social theory according to the ontological position adopted with respect to the micro-macro relationship. The majority of existing social theories, they argue, fall into one or other of the categories: *individualism* and *sociologism*. Neither of these 'paradigms' provides a theoretical foundation which supports exploration let alone the possibility of advancing understanding of the interplay between agency and structure. The third category, *dualism*, while considering both aspects, maintains a dichotomous stance as necessary and again does not advance any understanding of the interplay. Only those theories categorized as *dialectical* therefore have relevance. Even here, it is reasonable to conclude that little practical advance has been achieved, as most positions result in a straddling of bottom up and top-down arguments and/or suffer from excessively vague conceptualisation. These theories quickly break down again into a dichotomy the moment an attempt is made to make them operational.

What has been largely agreed, despite the very different theoretical and often inadequate handling of this problem, is that structure and agency come together in *activity* or in *body-hood* – the specific psycho-motor state at the instant of enaction. Both Vygotsky and Giddens, for example, focus on action as the point of intersection between human agency and social structures and it is implicit in Bourdieu's *habitus* also.

The Contribution from Systems Theory

Systems language was evident in the work of the early Emergentists and in sociology and anthropology which took seriously the structure/agency problem – notably that of Margaret Mead and Gregory Bateson. However, 'systems' as a focus of research took form with Bertalanffy's attempt to establish a General Systems Theory [28, 29]. As the science of 'wholes' systems theory stands in contrast to reductionisms

concern with parts: it was advanced as a counter to what was perceived as excessive reductionism dominating scientific discourse during much of the 20th century.

Early (first order) cybernetic approaches modelled systems as 'black boxes' effectively masking the relationship between micro and macro. Application of the concept to social science by Ernst von Glasersfeld and Heinz von Foerster [30] led to social (second order) cybernetics and soft systems approaches [31] more useful for describing the systemic behaviour of social systems. While the aspiration of the General Systems Movement to establish a general science of systems is widely regarded as having failed [32], systems approaches have contributed valuable methods for the study of the interplay between levels. The Systems view of emergence was founded on:

- Holism; the whole is greater than the sum of its parts.
- A concern with *positive and negative feedback*.
- A concern with boundaries and boundary conditions – including as an epistemic act rather than an ontological fact.

More recently the development of complex systems theory and its application to natural, social and cognitive phenomena has provided additional concepts upon which much current debate about emergence draws. Many of these concepts and methods have become widely used within the multi-agent modelling community [33-36].

In contrast to the position taken by the British Emergentists who argued that irreducibility was the *exception* [8], most real world systems are now argued to be non-linear [37-40] and hence irreducible. It is non-linearity which contributes to these system's capacity for novelty and unpredictability through the presence of deterministic Chaos [41, 42] and/or equifinality. Equifinality refers to a system where a single high level property may be realized by more than one set of micro-states which have no lawful relationship between them [12, 43, 44]. As there is no *a-priori* basis by which the likely micro state can be determined, such systems are irreducible and unpredictable in principle.

The Contribution of Theoretical Biology, Cognitive Science and Robotics

While complexity science has drawn on a diverse range of research threads, one area where an interest in emergent phenomena has been strongly represented is in Artificial Life [45] (Alife). While initially involving exploration of emergence using very simple 'cellular automata', there has been increased interest within this community to explain the fundamental building blocks of life. In contrast to first generation Artificial Intelligence [46] this has included a commitment to a bottom up methodology – i.e. evolving cognitive capability rather than engineering it in [47]. This has led the field to a biologically grounded perspective of cognition and one very different from the symbolic representation approach adopted within first generation AI. From this perspective any social emergent structures will be constrained by the biological fundamentals of cognition. In other words, behavioural and linguistic domains will depend on and be constrained by the metabolic systems which give rise to them. This has bridged Alife research into theoretical biology, in particular, autopoietic theory [47-49] and hence enactive theories of cognition [50, 51].

The enactive view of cognition was first proposed by the theoretical biologists Humberto Maturana and Francisco Varela [52, 53]. While these authors primary contribution has been towards understanding the self-organising metabolic mechanisms of life, the resulting theory of autopoietic systems provided a foundation for a general theory of cognition [54-56]. This *embodied/enactive* view stands in stark contrast to the symbolic representation [57], rational actor and game theoretical approaches which have most commonly informed social simulation. It has however recently seen considerable application in robotics [66-69], where it is argued to be fundamental to understanding how robots can become genuinely autonomous – i.e. capable of learning about their environment without the need for detailed information being provided by a designer. Within social theory some consideration has been given to the implications of enaction for understanding and theorising social behaviour [26, 58-60] although not without some controversy [61-63] and we have argued elsewhere that many of these extensions are incompatible with the original concept [64, 65]. None of this has yet found extension into social simulation.

Attempts to understand and specify mechanisms of social emergence have generally built upon the philosophical and systems theoretical literatures. There has been little accommodation of the wider debate about agency and structure particularly that associated with dialectical social theory. The micro level assumptions have been largely restricted to those associated with the rational actor and game theory and first generation AI. Very little work has been done to incorporate the perspective offered by recent developments in artificial life, robotics and theoretical biology. It is however this detailed work on the relationship between cognitive capability and associated emergent behaviour that arguably provides the most valuable contribution to our understanding of social emergence. This is in part due to it being grounded in the study of real biological entities and/or the practical challenges of building viable robots.

3 ORDERS OF EMERGENCE

A number of authors have identified what they refer to as orders of emergence. Gilbert, for example distinguishes between a first and second order. First order emergence includes macro structures which arise from local interactions between agents of limited cognitive range (particles, fluids, reflex action). By contrast, second order emergence is argued to arise *'where agents recognise emergent phenomena, such as societies, clubs, formal organizations, institutions, localities and so on where the fact that you are a member or a non-member, changes the rules of interaction between you and other agents.'* [70]. This reflects high order cognition – in particular a capacity to distinguish class characteristics, assess 'self' for conformity and to change behaviour accordingly. First and second order emergence then each imply qualitatively distinct cognitive mechanisms and suggest a continuum of orders of emergence linked to cognitive capability.

In a similar vein, Castelfranchi [2: 27] has distinguished 'cognitive emergence' which: *'... occurs where agents become aware, through a given 'conceptualization' of a certain 'objective' pre-cognitive (unknown and non deliberated) phenomenon that is influencing their results and outcomes, and then, indirectly, their actions.'* This approach is based on a first generation AI [46] approach to conceptualizing agents – agent

cognition is assumed to involve acting on beliefs desires and intentions (BDI). Thus Castelfranchi conceives of a feedback path from macro pattern to micro behaviour in much the same way as Gilbert, except that here a cognitive mechanism is specified. Castelfranchi argues that this mechanism *'characterises the theory of social dynamics'* and gives rise to a distinct class of emergent phenomena. In this account, the representations agents have about the beliefs, desires and intentions of other agents plays a causal role in their subsequent behaviour and therefore shapes the structures they participate in generating. Castelfranchi argues that understanding this process is fundamental to social simulation: it is where social simulation can make its greatest contribution.

These ideas are more comprehensively reflected in the five orders of emergence suggested by Ellis [71:99-101].

Order	Ellis' Description of Properties	Characteristic Organization[3]
1	Bottom up leading to higher level generic properties (examples include the properties of gases, liquids and solids)	Property emergence
2	Bottom up action plus boundary conditions lead to higher level structures (e.g. convection cells, sand piles, cellular automata)	Self-organization Far-from-Equilibrium (weak-autonomy)
3	Bottom up action leading to feedback and control at various levels leading to meaningful top down action - teleonomy (e.g. living cells, multi-cellular organisms with 'instinctive' – phylogenetically determined reactive capability)	Self-production (autopoiesis) of metabolism (strong autonomy)
4	as per 3 but with the addition of explicit goals related to memory influence by specific events in the individual history (i.e. capable of learning)	Autonomous sensori-motor loops (strong autonomy)
5	In addition to 4 some goals are explicitly expressed in language (humans).	Semiotic autonomy (strong autonomy)

Table one: Adapted from Ellis.

In table one we set out Ellis' order number in column one and his description of the associated characteristics in column two. In column three we suggest an alternative classification which draws on the distinctions suggested by Rocha.

As with the approach of Gilbert and Castelfranchi, Ellis's framework also suggests that the range and type of emergence possible depends fundamentally on the range and class of behaviour agents are able to generate.

Considering category one emergence: particles have fixed properties and are able to enter into a limited range of interactions with others based on those properties. Swarms of particles can nevertheless demonstrate some rudimentary self-organisation and hence emergence [45]. Physics has furnished good accounts of many specific examples [73] but they have limited implication for our understanding of social behaviour.

Category two has also been well explored – it is the focus of a great deal of the work undertaken on complex, far from equilibrium systems [74, 75]. Examples include the work of Per Bak [76] on sand piles and earthquakes and Lorenz [42] on weather systems. Many so called social simulations which incorporate agents which have fixed behaviours and no capacity for learning also arguably belong here. These include classic simulations such as Schellings segregation model, the cooperation models of Axelrod [77] or the Sugarscape models of Epstein and Axtell [78]. Some may argue that these models involve agents with goals and therefore represent examples of fourth order emergence. The transition between third order and fourth, as will be argued below, involves a move to agent autonomy that is missing in these models – their goals are designed in and not a result of their own operation – it is for this reason that we argue they belong to order two although some may argue they represent reasonable analogues of the type of behaviour that might be generated by agents with higher order capability.

It is significant that Ellis' provides primarily biological examples for his category three order of emergence. The paradigmatic biological entity which illustrates the processes of reciprocal micro-macro causality pointed to by Ellis and for which we have an excellent description which has been made operational in vitro and in silico [see for example 79, 80] is the cell. While the mechanisms of autocatalysis and the metabolic pathways of cell self-production are well documented and closely studied, the most concise articulation of the fundamental self-producing processes involved comes in the theory of autopoiesis already mentioned [52, 53, 80, 81]. Varela [82: 78] states: *Autopoiesis is a prime example of a …dialectics between the local component levels and the global whole, linked together in reciprocal relation through the requirement of constitution of an entity that self-separates from its background.* In other words the distinguishing characteristic in this order is that the micro-macro interplay leads to an autonomous structure which acts so as to maintain its viability as an entity. This is not the case for many far from equilibrium systems such as weather systems. The maintenance of viability is a clear threshold and one we appear far from being able to simulate using existing methods.

In his third order category Ellis includes a range of capabilities of biological entities up to and including 'instinctive' action. These suggest that single and multi-cellular organisms including those with a central nervous system would all be included. It may be that this order is too broadly cast. Ellis has grouped entities such as cells which rely exclusively on metabolic self-regulation with entities which also have a capacity to self-regulate using sensorimotor mechanisms. Differentiated aggregates of cells display greater capacity to respond to their environment, even where they do not possess a central nervous system, than do individual cells (e.g. by development of an immune response). A central nervous system provides the entity with even greater behavioural plasticity [52] and hence a capacity to maintain its viability in a wider range of environments. As a consequence each threshold probably originates a distinct macro phenomenology different from that of the cells that constitute them [53].

The primary point of distinction between order three and order four would appear to be between (phylogenetically) fixed individual characteristics and a capacity for an individual agent to learn. This category covers animals up to human but this again is a big span covering a number of cognitive and developmental thresholds, including the emergence of pre-linguistic theory of mind, and self-awareness [83] which might be expected to have

[3] Classification according to the work of [72. Rocha, L.M., *Language Theory: Consensual Selection of Dynamics.* Cybernetics and Systems, 1996. **27**(6): p. 541-553.]

a significant effect on social emergence. It is also not clear what is meant by learning. Learning can span a wide range of capabilities from simple operant conditioning to advanced reasoning.

The final transition between order four and five demarcates the line between non-human animals and humans. The advent of language gives rise to a distinct phenomenal domain with significant implications for social emergence. This is not least due to the association language has in humans with other cognitive capabilities such as theory of mind, narrative ability and reflexivity.

Examining the characteristic organization implied in Ellis' orders of emergence shows that the transition points are strongly linked to processes of self-organisation and autonomous closure. Furthermore this autonomous closure occurs recursively: closure at one level makes possible closure at a higher level and so on. What we are essentially attempting to do in social simulation at present is to shortcut this process: to achieve reasonable analogues of behaviour at various levels without also modelling the processes upon which it depends. This appears reasonable – we do not need to model sub-atomic processes in order to work with models of molecules and understand the reaction chains they can participate in, so why would we need to model metabolic or sensorimotor systems in order to understand social interactions? How then do we advance our understanding of the effect of different cognitive capability on orders of emergence and if and when they matter?

4 AGENT AUTONOMY

Robots are generally intended to be able to perform useful functions in real and complex environments. To do so they need to have a level of autonomy: a capacity to map their worlds and to decide what is important and change their behaviour accordingly. This proved computationally difficult (if not impossible) to achieve using conventional AI approaches. A breakthrough was achieved with Brooks demonstration of the power of situated cognition [84]. It is therefore no surprise that our understanding of the implications and opportunities presented by understanding cognitive autonomy has been led by the field of robotics. What then is the state of the art and what implications may it have for understanding and simulating social emergence?

In her introductory paper for the Modelling Autonomy Workshop held in San Sebastián in March 2007 (http://www.ehu.es/ias-research/autonomy/), Margaret Boden stated that *'very broadly speaking, autonomy is self-determination: the ability to do what one does independently, without being forced so to do by some outside power.'* She notes that the concept is problematic as there are various types and degrees of independence. This has already been illustrated above when examining Ellis' orders of cognition. In social simulation we have achieved limited independence in the form of self-organization. For Barandiaran & Moreno [85: 179], *'The main difference between self-organization and autonomy is that while self-organization appears when the (microscopic) activity of a system generates at least a single (macroscopic) constraint, autonomy implies an open process of self-determination where an increasing number of constraints are self-generated.'* This reemphasises that autonomy involves recursion: cyclic generation proceeding from simple self-organisation to closure in a succession of phenomenal domains culminating in closure at the semiotic level.

Within Alife and robotics, it has been increasingly argued that while autopoiesis specifies the metabolic closure and self-production characteristic of living entities, cognition implies more than this. A cognitive agent has a primary autonomous metabolic loop which serves to maintain its biological viability and (at least) one other loop which links sensory surfaces with motor surfaces [see also 49]. This second loop affords the agent significant additional plasticity. This plasticity is realised within a behavioural rather than a metabolic phenomenal domain [86: 168]. The two are interdependent in that the range of behaviour the agent can generate is dependent on its biology, while its biological viability can depend on the behaviour: the recognition and escape from threat or the location of food for example. While Duijn et al argue that this sensorimotor loop is already present in the two component signal transduction system (TCST) system found in bacteria, Moreno et al [47] argue that it is the central nervous system which fundamentally distinguishes biological/metabolic processes from cognitive processes.

Irrespective of where this line is drawn, both are consistent in the view that *'…cognition is not so much a centralized property of the biological hardware of an organism, or a set of internally computed algorithms, but instead denotes an abstraction of organism environment reciprocity'*.[87]. This is consistent with the position taken by Varela [50, 82, 88], that autonomous agents 'bring forth a world' as a result of their operational closure. In other words, what an agent can perceive and cognize is determined by its own operation, not the environment. Again from Barandiaran, under conditions of autonomy: *'It is not the organism that matches the environment in a given specified way. On the contrary it is through the particular way in which the agent satisfies the homeostatic maintenance of essential variables that an adaptive environment (a world) is specified - cut out from a background of unspecific physical surroundings.' [49]*

What this means is that the environment is a source of perturbations which act only as triggers for change. It is the nervous system's structure that dictates which perturbations can be a trigger [57, 89]. Consequently changes to the structure of one agent's nervous system, and consequently its behaviour, will be unique to that agent. The environmental perturbations that act as a change trigger in one agent will not necessarily trigger a change in another, or if they do, the change that is triggered may take a different form and/or have different implications for the viability of that agent in its environment, given its history of interactions.

The consequence of this recursive construction of increasing order of autonomy for the agent is enhanced viability in a wider range of environments. This is apparent if we consider the effect of the transition from metabolic autonomy (autopoiesis) to sensorimotor autonomy supported by a central nervous system. The coexistence of these two interdependent levels of autonomous functioning allows the organism to exploit the rapid response times of the neural system and this makes possible a significantly increased set of possible responses to environmental perturbations [49]. An organism that relies less on the slow diffusion reactions associated with metabolism, and which can draw on the rapid response of the chemical/electrical nervous system is better able to survive in less stable

environments. In systems terms, it has greater requisite variety [90].

It is this asymmetry between the state space of possible configurations made possible by an advanced nervous system and the range of response needed to maintain immediate regulation that gives rise to what we call 'agency'. Agency is a consequence of autonomy. Agency makes possible what we typically regard as distinguishing features of social systems: endogenous goal making and seeking behaviour. Agency also supports 'free-will', the opportunity for agents to behave in ways which are non-deterministic: to generate new bottom up solutions to situations they encounter. From this perspective then autonomy is fundamental to agency and hence to a capacity to engage in activity which can genuinely be called social. We conclude therefore that to be able to simulate an agent which is deserving of the title of being a social agent, it would need to exhibit some level of 'strong autonomy' and hence agency. But we are a long way from achieving it.

Why is this necessary and what does it reveal about the fundamental mechanisms at work in social systems? Also, at this stage we have considered only agents in isolation. What happens when we bring multiple autonomous agents together such that they can interact?

5 MECHANISMS OF SOCIALITY

Following the line of argument developed above, when brought together each agent treats each other agent as a part of its environment.

As agents interact each undergoes a set of internal structural accommodations which allow it to persist in its relationship with the others. This results in a 'structural drift', or a gradual change to the state of each agents nervous system [52, 53]. Over time it traces a unique history − Maturana refers to this as the agent's *ontogeny*. When interactions become 'recurrent' – that is repetitive and ongoing – agents can become 'structurally coupled'. Here we have the most basic element of sociality and one that can be applied to all organisms with nervous systems, even very elementary ones.

Importantly, a history of recurrent interactions leads to a structural congruence or commonality of experience between two or more agents: their behaviours become tuned to one another in a reciprocal 'dance' maintained in and through their relating. The degree of structural coupling that arises when two or more agents interact is a fundamental factor in determining the dynamics and emergent behaviour of the resulting structurally coupled system. Agents give rise to a behavioural phenomenal domain in which a range of attractors may form. These attractors are what we would typically call macro-social structures. What then of the advent of language?

Language is associated with higher order cognition. It will support a behavioural domain which is more inherently plastic than one coordinated only through bodily interaction. Otherwise, as a mechanism, it is an extension of what has already been discussed. It is however a non trivial extension as the state space of points of interaction becomes very much larger where the variables (utterances) are recursive as they are in language as a) agents make linguistic distinctions on linguistic distinctions b) new linguistic constructs are under control of the system they also serve to regulate.

Structural coupling within a linguistic domain will be apparent from the convergence of the individual linguistic utterances to form a shared lexicon and grammar. The driving force behind this convergence is the one fixed internal goal the agents have: that of maintaining their viability. If they are biological agents this will involve the preservation of their autopoiesis – i.e. remaining alive. At base level this involves meeting the requirements of the metabolic level of operation. They will need to eat, stay warm, avoid predators and find partners. The metabolic or biological necessarily interacts with the behavioural and the linguistic domains: the domains are co-dependent.

The agents will innovate in their behaviour in order to satisfy their minimal requirements. Some of the behaviours they adopt will, however, be due to a need to accommodate the behaviours of other agents. A set of attractors should therefore emerge which represent sets of states which 'satisfice' social constraints as well as fundamental biological constraints. To an observer[4] some of these states may appear as goal based (food seeking) while others may be seen to be primarily to do with mutual accommodation (norms). The attractors may be reflected in macro structuration (division of labour, identity groups) and may assemble into yet higher order patterns (organizations, institutions). The engine of this process of social emergence is structural coupling and the dimensions of possible coupling and the scope of behaviours which may be involved in establishing and maintaining coupling is dependent on the biology and cognitive plasticity of the agent.

We are currently able to simulate behaviours up to order three. Moving beyond this raises some interesting questions. Among these are: are the cognitive capabilities clearly associated with social behaviour necessarily tied to metabolic autonomy? To what degree do these capabilities manifest the way they do due to the specific organic mechanisms associated with life? Is it possible to simulate behavioural autonomy and linguistic systems which are operationally closed on other than an organic substrate? If so what are the essential low level characteristics which are essential to supporting them? In short: is it possible to model these types of processes in-silico?

6 CONCLUSION AND FUTURE DIRECTIONS

There is a range of ways of thinking about the relationship between micro and macro level phenomena. There have been centuries of debate about the relative merit of reductionist, vitalist and holistic perspectives for understanding how higher order structures emerge from lower. Despite ongoing scepticism in some philosophical quarters, we have advanced our understanding of the mechanisms involved to a very significant degree over the past 30 years. Emergent structures are increasingly understood to be a product of non-linear interactions associated with complex systems of agents. We can however, go further than this. One of the insights being

[4] The changing role of the observer in distinguishing different categories of emergence is a significant issue which has not been taken up here. Arguably observer based distinctions play no role in the property emergence associated with Ellis' category one, whilst they are intrinsic to the social emergence associated with category five. The role in the intermediate categories is less clear. We have set out some early thinking about the critical role this plays in 91. Goldspink, C. and R. Kay. *Social Emergence: Distinguishing Reflexive and Non-reflexive Modes* in *AAAI Fall Symposium* 2007. Washington.

developed is that the range and type of emergent structure depends on the specific mechanisms involved and on the properties of the micro agents.

This paper has concentrated on how alternative micro-capabilities support qualitatively distinct forms of social emergence. What has been argued is that social emergence implies not a single transition from micro to macro but is built upon, and is an example of recursive self-organization within a biological domain. The recursive levels in living systems span metabolic, neurological, social-behavioural and social-semiotic levels.

Social emergence involves a level of self-reference and self generation which is not apparent in non-organic forms of far from equilibrium behaviour. Social emergence builds on biological emergence. This is to say that the phenomenal domains associated with social systems, particularly those involving humans, are constrained by the biological processes which make them possible. The ongoing debate about emergence as a concept demonstrates that understanding the relationship between micro and macro phenomena is theoretically as well as practically challenging. To date it has proven difficult to build models which provide reasonable analogues of this process. What has been achieved has been achieved largely in robotics, and Artificial Life. So far, social simulation has played a minor role. Nevertheless the science of these processes is important to social simulation. It has proven possible to model some social behaviour to good effect without agents with these capabilities. It can also be argued that in highly complex (chaotic or random) environments higher order cognition is of little value justifying a parsimonious substitution of particle-like agents. However, it is reasonable to expect that we will not be able to effectively model some forms of social behaviour without having come to terms with and found ways to simulate behaviour which is possible due to autonomous closure. Equally social simulation could play an important role in helping us to understand the implications of autonomous closure and for advancing our ability to theorise about it.

References

1. Sawyer, K.R., *Artificial Societies: Multiagent Systems and the Micro-macro Link in Sociological Theory.* Sociological Methods & Research, 2003. **31**: p. 38.
2. Castelfranchi, C., *Simulating with Cognitive Agents: The Importance of cognitive emergence*, in *Multi-agent Systems and Agent Based SImulation*, J.S. Sichman, R. Conte, and N. Gilbert, Editors. 1998, Springer: Berlin.
3. Goldspink, C. and R. Kay, *Organizations as Self Organizing and Sustaining Systems: A Complex and Autopoietic Systems Perspective.* International Journal General Systems, 2003. **32**(5): p. 459-474.
4. Goldspink, C. and R. Kay, *Bridging the Micro-Macro Divide: a new basis for social science.* Human Relations, 2004. **57** (5): p. 597-618.
5. Ablowitz, R., *The Theory of Emergence.* Philosophy of Science, 1939. **6**(1): p. 16.
6. Peterson, G.R., *Species of Emergence.* Zygon, 2006. **41**(3): p. 22.
7. Shrader, W.E., *The Metapysics of Ontological Emergence*, in *Graduate School*. 2005, University of Notre Dame.
8. Eronen, M., *Emergence in the Philosophy of Mind*, in *Department of Philosophy*. 2004, University of Helsinki: Helsinki. p. 79.
9. Stanford Encyclopaedia of Philosophy, *Emergent Properties*, in *Stanford Encyclopaedia of Philosophy*. 2006.
10. Clayton, P. and P. Davies, *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*. 2006, Oxford: Oxford University Press.
11. Fodor, J.A., *Special; Sciences or The Disunity of Science as a Working Hypothesis.* Synthese, 1974. **28**: p. 18.
12. Sawyer, K.R., *Emergence in Sociology: Contemporary Philosophy of Mind and Some Implications for Sociology Theory.* American Journal of Sociology, 2001. **107**(3): p. 551-585.
13. Symons, J., *Emergence and Reflexive Downward Causation.* Principia, 2002. **1**: p. 183-202.
14. Davies, P., *The Physics of Downward Causation*, in *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, P. Clayton and P. Davies, Editors. 2006, Oxford University Press: Oxford.
15. Engels, F., *Dialectics of Nature*. 1934, Moscow: Progress Publishers.
16. Vygotsky, L.S., *Thought and Language*, ed. T.b.E.H.a.G. Vakar. 1962, Cambridge, Mass: MIT Press.
17. Leont'ev, A.N., *Activity, Consciousness and Personality*. 1978, Engelwood Cliffs: Prentice Hall.
18. Berger, P.L. and T. Luckman, *The Social Construction of Reality*. 1972: Penguin.
19. Giddens, A., *The Constitution of society: Outline of the theory of structuration*. 1984, Berkeley: University of California Press.
20. Bhaskar, R., *A Realist Theory of Science*. 1997, Verso: London.
21. Bhaskar, R., *The Possibility of Naturalism*. 1998, London: Routledge.
22. Archer, M., *Realism in the Social Sciences*, in *Critical Realism: Essential Readings*, M. Archer, et al., Editors. 1998, Routledge: London.
23. Archer, M., et al., *Critical Realism: Essential Readings*. 1998, London: Routledge.
24. Coleman, J.S., *Foundations of Social Theory*. 1994, Cambridge: Belknap.
25. Luhmann, N., *Essays on Self Reference*. 1990, New York: Columbia University Press.
26. Luhmann, N., *Social Systems*. 1995, Stanford: Stanford University Press.
27. Fuchs, C. and W. Hofkirchner, *The Dialectic of Bottom-up and Top-down Emergence in Social Systems.* tripleC 2005. **1**(1): p. 22.
28. Bertalanffy, L.v., *An Outline of General Systems Theory.* British Journal for the Philosophy of Science, 1950. **1**(2).
29. Bertalanffy_von, L., *General Systems Theory*. 1968, New York: Braziller.
30. Keeney, B.P., *Aesthetics of Change*. 1987: Guilford.
31. Checkland, P., *Systems Thinking Systems Practice*. 1988, G.B.: John Wiley.
32. Jackson, M.C., *Systems Approaches to Management*. 2000, London: Kluwer Academic.
33. Holland, J.H., *Emergence: from chaos to order*. 1998, Ma.: Addison Wesley.

34. Gilbert, N., *Emergence in Social Simulation*, in *Artificial Societies*, N. Gilbert and R. Conte, Editors. 1995, UCL Press: London.
35. Castelfranchi, C., *Simulating with Cognitive Agents: The Importance of Cognitive Emergence*, in *Lecture Notes in Artificial Intelligence*, J.S. Sichman, R. Conte, and N. Gilbert, Editors. 1998, Springer Verlag: Berlin.
36. Conte, R., R. Hegselmann, and P. Terna, *Simulating Social Phenomena*. 1997, Berlin: Springer.
37. Stewart, I., *Does God Play Dice - The New Mathematics of Chaos*. 1990: Penguin.
38. Kauffman, S., *Investigations*. 2000, New York: Oxford.
39. Kauffman, S.A., *The Origins of Order: Self Organization and Selection in Evolution*. 1993: Oxford University Press.
40. Kauffman, S.A., *At home in the Universe: The Search for Laws of Complexity*. 1996, London: Penguin.
41. Williams, G.P., *Chaos Theory Tamed*. 1997, Washington D.C: Joseph Henry Press.
42. Lorenz, E.N., *The Essence of Chaos*. 4 ed. 2001, Seattle: University of Washington Press.
43. Richardson, K.A., *Methodological Implications of a Complex Systems Approach to Sociality: Some further remarks.* Journal of Artificial Societies and Social Simulation, 2002. **5**(2).
44. Richardson, K.A. *On the Limits of Bottom Up Computer Simulation: Towards a Non-linear Modeling Culture*. in *36th Hawaii International Conference on Systems Science*. 2002. Hawaii: IEEE.
45. Kennedy, J. and R.C. Eberhart, *Swarm Intelligence*. 1 ed. 2001, London: Academic Press. 511.
46. Franklin, S., *Artificial Minds*. 1998, London: MIT press.
47. Moreno, A., J. Umerez, and J. Ibanes, *Cognition and Life.* Brain and Cognition, 1997. **34**: p. 107-129.
48. Beer, R.D., *Autopoiesis and Cognition and the Game of Life.* Artificial Life, 2004. **10**: p. 309-326.
49. Barandiaran, X., *Behavioral Adaptive Autonomy. A Milestone on the ALife roue to AI?* 2005, Department of Logic and Philosophy of Science, University of the Basque Country: San-sebastian, Spain.
50. Rudrauf, D., et al., *From Autopoiesis to Neurophenomenology: Francisco Varela's exploration of the biophysics of being.* Biol. Res, 2003. **36**: p. 27-65.
51. Thompson, E. and F.J. Varela, *Radical embodiment: neural dynamics and consciousness.* TRENDS in Cognitive Sciences 2001. **5**(10): p. 7.
52. Maturana, H. and F. Varela, *Autopoiesis and Cognition: The Realization of the Living*. Boston Studies in the Philospohy of Science. Vol. 42. 1980, Boston: D. Reidel.
53. Maturana, H.R. and F.J. Varela, *The Tree of Knowledge: The Biological Roots of Human Understanding*. 1992, Boston: Shambhala.
54. Maturana, H., *Ontology of Observing: The Biological Foundations of Self-Consciousness and of The Physical Domain of Existence*. 1988.
55. Maturana, H., *The origin and conservation of self-consciousness Reflections on four questions by Heinz von Foerster.* Kybernetes 2005 **34**(1/2): p. 34.
56. Maturana, H., J. Mpodozis , and J.C. Letelier, *Brain, Language and the Origin of Human Mental Functions.* Biological Research 1995. **28**: p. 11.
57. Varela, F., E. Thompson, and E. Rosch, *The Embodied Mind*. 1992, Cambridge: MIT Press.
58. Teubner, G., *How the Law Thinks: Towards a Constructivist Epistemology of Law.* Law and Society Review, 1989. **23**(5): p. 727-758.
59. Zeleny, M., *Autopoiesis: A Theory of Living Organization*. 1991, New York: North Holland.
60. von_Krogh, G. and J. Roos, *Organizational Epistemology*. 1995, London: St Martins Press.
61. Mingers, J., *Are Social Systems Autopoietic? Assessing Luhmanns Social Theory.* Sociological review, 2002. **50**(2).
62. Mingers, J., *Can Social Systems be Autopoietic? Bhaskar's and Giddens' Social Theories.* Journal for the Theory of Social Behaviour, 2004. **34**(4): p. 25.
63. Bednarz, J., *Autopoesis: the Organizational Closure of Social Systems.* Systems Research, 1988. **5**(1): p. 57-64.
64. Goldspink, C., *Social Attractors: An Examination of the Applicability of Complexity theory to Social and Organisational Analysis*, in *Social Ecology*. 2000, University Western Sydney: Richmond. p. 312.
65. Kay, R., *Towards an autopoietic perspective on knowledge and organisation*, in *Social Ecology*. 1999, University of Western Sydney: Richmond.
66. Di Paolo, E.A., *Autopoiesis, adaptivity, teleology, agency*. n.d, Department of Informatics, University of Sussex: Brighton.
67. Di Paolo, E.A., *Organismically-inspired robotics: homeostatic adaptation and teleology beyond the closed sensorimotor loop*. n.d, School of Cognitive and Computing Sciences, University of Sussex: Brighton.
68. Di Paolo, E.A. and H. Lizuka, *How (not) to Model Autonomous Behaviour.* Biosystems, 2007.
69. Di Paolo, E.A., M. Rohde, and H. De Jaegher, *Horizons for The Enactive Mind: Values, Social Interaction and Play*, in *Enaction: Towards a New Paradigm for Cognitive Science*, J. Stewart, O. Gapenne, and E.A. Di Paolo, Editors. 2007, MIT Press: Cambridge MA.
70. Gilbert, N. *Varieties of Emergence*. in *Social Agents: Ecology, Exchange, and Evolution Conference* 2002. Chicago.
71. Ellis, G.F.R., *On the Nature of Emergent Reality*, in *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, P. Clayton and P. Davies, Editors. 2006, Oxford University Press: Oxford.
72. Rocha, L.M., *Language Theory: Consensual Selection of Dynamics.* Cybernetics and Systems, 1996. **27**(6): p. 541-553.
73. Gell-Mann, M., *The Quark and the Jaguar: Adventures in the simple and the complex*. 1995, Great Britain: Abacus.
74. Prigogine, I., *The End of Certainty: Time, Chaos and the New Laws of Nature*. 1997, New York: The Free Press.
75. Prigogine, I. and I. Stengers, *Order out of Chaos: Man's New Dialogue with Nature*. 1985: Flamingo.
76. Bak, P., *How Nature Works: The Science of Self-Organized criticality*. 1996, New York: Copurnicus.
77. Axelrod, R., *The Evolution of Cooperation*. 1984, New York: Basic Books.
78. Epstein, J.M. and R. Axtel, *Growing Artificial Societies*. 1996, Cambridge, Ma.: MIT Press.
79. McMullin, B. and D. Grob, *Towards the Implementation of Evolving Autopoietic Artificial Agents*, in *6th European Conference on Artificial Life ECAL 2001*. 2001: University of Economics, Prague.

80. Varela, F., H. Maturana, and R. Uribe, *Autopoiesis: The Organization of Living Systems, Its Characterization and a Model.* Biosystems, 1974. **5**: p. 187-196.
81. Varela, F., *Principles of Biological Autonomy*. 1979, New York: Elsevier-North Holland.
82. Varela, F., *Patterns of Life: Intertwining Identity and Cognition.* Brain and Cognition, 1997. **34**: p. 72-87.
83. Gardenfors, P., *How Homo became Sapiens: On the evolution of Thinking*. 2006, Oxford: Oxford University Press.
84. Brooks, R.A., *Intelligence without representation.* Intelligence without reason, 1991(47): p. 569-595.
85. Barandiaran, X. and A. Moreno, *On what makes certain dynamical systems cognitive: A minimally cognitive organization program.* Adaptive Behavior, 2006. **14**(2): p. 171-185.
86. Moreno, A. and A. Etxeberria, *Agency in natural and artificial systems*. 1995, Department of Logic and Philosophy of Science University of the Basque Country: San Sabastian, Spain.
87. Duijn, M.v., F. Keijzer, and D. Franken, *Principles of Minimal Cognition: Casting Cognition as Sensorimotor Coordination.* Adaptive Behavior, 2006. **14**(2): p. 157–170.
88. Thompson, E., *Life and Mind: From Autopoiesis to neurophenomenology, a Tribute to Francisco Varela.* Phenomenology and the cognitive Sciences, 2004. **3**: p. 381-398.
89. Mingers, J., *The Cognitive Theories of Maturana and Varela.* Systems Practice, 1991. **4**(4): p. 319-338.
90. Ashby, W.R., *Self-regulation and Requisite Variety*, in *Sytems Thinking*, F.E. Emery, Editor. 1974, Penguin: Great Britain.
91. Goldspink, C. and R. Kay. *Social Emergence: Distinguishing Reflexive and Non-reflexive Modes* in *AAAI Fall Symposium* 2007. Washington.

Formalizing Epistemological Constituents of Emergence

Raif Serkan Albayrak[1] and **Ahmet Süerdem**[2]

Abstract. Social action depends on the knowledge about all levels of emergent structure and this knowledge is not limited to but produced by the local binary interactions, thus requiring reflexivity and intentionality. This knowledge resides in the impersonal area of symbol systems and is produced by a meta-language which carries the knowledge of emergent structure beyond local networks to the topological knowledge of the macro phenomena. Individual agents use symbol systems as tool-kits to encode their intents, and use their world-views to ground the symbols and ethos to challenge meaning-symbol correspondences through their everyday practices. Symbolic interaction is not only the symbolic affirmation of shared social classifications and normative protocols that regulate interactions but is also making sense of expressive, symbolic behaviour and decoding the intent of the counterparts from these symbols. Meta-language is the key concept to understand how this knowledge is generated through semiotic relations. This paper develops the formal infrastructure of such a model and elaborates various mechanisms that can be implemented within a social simulation model.

1 INTRODUCTION

The controversy between emergence of distinct social phenomena through interactions of individual agents and social causation by macrostructures has been a major issue in social sciences. This debate about the effects of agency and structure on human cognitive abilities and behaviours has recently been carried to social simulation community and challenged its basically agent based character. The debate up to now stayed at the ontological level. Discussions are carried on around the nature and origins of social phenomena (reductionism-holism); the operation of mechanisms of emergence (bottom-up-top-town); and the properties of the cognitive abilities of the agents (reactive-deliberative).

However, not much has been said about the epistemological constituents of emergence. How individual agents come to know the mechanisms of emergence is a relatively less referred topic. In this vein, this paper aims to contribute to the debate by formalizing a meta-language approach which explains the mechanisms of how individual agents can decode and feed-back accumulated knowledge about the emerging macrostructure.

In ABSS models developed so far, social agents have knowledge about the cognitive capabilities and properties of specific other agents although they are not given the possibility to reach knowledge about emergent macro phenomena. These models follow a reductionist approach: they start by modelling built-in cognitive devices to individual agents; determine the rules of interaction among them; and hence try to observe emergent macro patterns that influence their behaviour and interactions [9]. However, within this framework, emergence is only possible if the interconnection between the agents is functionally determined. Agents need predefined codes to make their and their counterparts' actions predictable and hence to feedback to the emergent phenomena.

This approach becomes problematic when the functionality of the interconnections assumption is violated. In case of complexity, where the relations between the components are from one to many and non-linear, mechanisms of system cannot be reduced to the properties of its parts. Information from emerging macro structure becomes an independent variable itself [31]. Hence, interconnections between components become fuzzily reciprocal and this makes it difficult to determine any reducible functional rule.

However, real world is complex and emergence is a real life phenomena. Within these premises, we either have to give up complexity for the sake of modelling or give up emergence for the sake of irreducibility when we are into explaining the relation between many levels of society. We believe that recent debates within the agent community present a sound platform to solve this critical issue [30,31,16,7,9,19]. Furthermore these efforts to bridge the micro-macro divide within the agent community have promising counterparts in the AI (Artificial Intelligence) community who redefine cognition as an embedded and embodied activity emerging from the dynamic interaction between brain, body, and environment [1,4,14]. In our view, theory of embodied cognition (EC) and a theory of sociality founded on a synthesis of autopoietic and complexity theories appears to be a strong candidate for solving this paradox [19].

Autopoietic theory determines the rules supporting the maintenance of self-production of organisms as operationally closed systems where the brain is not the coordinator but a part of the nervous system [19,20,2]. Organism's behaviour is more than a response to stimuli, it is a function of past modifications to the nervous system and thus unique to every individual. Goldspink and Kay [19; p. 603] name the unique history defined by each individual's history of interactions with the environment as *ontogeny*. This concept is strikingly parallel to "the lived experience" (*le vecu*) concept in phenomenology, and gives us the opportunity to introduce EC to autopoietic and complexity theories. According to phenomenologist, the phenomenal is not an object out there but is constructed through our bodily and sensory functions. One of the prominent figures in phenomenology, Merleau-Ponty, stated that intentional objects of thought (*noumenal*) cannot be separated from the perceived objects by thought (*phenomenal*). Cognition is embodied and this embodiment resides in the unique history or "the lived" (*le vecu*), experiences of the body. [29]

People do not live in a social vacuum and positioning within the social space is the essential lived experience for the human body. When two or more people interact, their lived experiences will mutually modify each other and their system will embody

[1] Yasar University, Turkey, email: raif.albayrak@yasar.edu.tr

[2] Bilgi University, Turkey, email: asuerdem@bilgi.edu.tr

the perturbations created by the emerging interactions. *Ontogeny* emerges within a consensual domain between the individuals who have lived similar experiences. When these interactions start to constitute repetitive patterns, individuals become "structurally coupled" thus are embedded within the emerging structure [19]. Once structurally coupled, common experiences start to become the common "reality" or "life world" [29] of the interacting individuals. Emergent life-world constitutes a base for social coordination and emotional and cognitive patterns generated by the common experiences orient individuals to construct identities, coordinate action, and create cooperation [33]. Life-world is the common *leitmotiv* underlying capabilities, practices and behaviours residing cognitive repertoire of the individuals that form a community [21].

Although marrying autopoiesis and EC provides us with a useful tool to understand the interrelation between cognition and emergent social phenomena, it still lacks explanatory power to understand how phenomena at distinct levels of the system are connected to each other. Especially, when structurally coupled life-worlds are self-referential and operate strictly according to their very own codes and have no knowledge of how the others decode their environment, we can hardly understand how social order is maintained at the macro-level.

The key element to solve this problem is a theory of communication since social systems are systems of communication [28]. An autopoietic system has well defined boundaries between itself and an infinitely complex environment. Communication within itself operates according to selection of only a limited amount of all information available outside. This helps the system to reduce complexity. The criterion according to which information is selected and processed depends on meaning production: complexity reduction is generating patterns that can recognize the environment and couple with the phenomena in a coherent way.

According Goldspink and Kay [19] human capacity for language is the key to model fuzzy and non-functional relations between autopoietic agents. Contrary to natural systems where the behaviour of the individuals are activated by local influences only, social systems can handle the problem of complexity through a feedback mechanism which allows changes induced in the macrostructures to be felt locally. This feedback mechanism occurs by means of linguistic activities that provide agents with reflexive capacities. Such capacities endow agents to decode macro-patterns and encode their local behaviour accordingly. Language provides them with a foundation for a flexible and instant feedback about the macro-structure.

Yet, communicative practices should not be reduced to a grammar of coded equivalences as posited by the linguistic semiotics. Such a coding system will not be suitable for modelling complex dynamics. Autopoietic, complexity and EC theories link the biology of cognition to the nature of the human linguistic capacity as rooted in the dynamics of reciprocal causality between an organism and the environment. Accordingly, a dynamic model of language cannot be denotational but needs to be constructed on connotational principles [26,27]. Such a model of meaning construction requires an active, generative process driven by one-to-many symbol oppositions rather than a passive mapping of mental representations. Meta-language is the key concept to understand how meaning is generated through opposition between symbols and in the rest of the paper we will present its mathematical formalization.

2 META-LANGUAGE

The basic element of meta-language is symbols and we assume that they exist and they are everywhere. However in order to realize their existence they must be distinguishable. We argue that every symbol reveals its identity from its distinctions to other elements in the same system. In a roughly Derridean way [35], we will name such distinctions as oppositions. In order to refine our stand from the controversy on the constituents of an opposition [36] we note that our treatment of an opposition stands for a distinguishing character in terms of connotations. Derridean deconstruction assumes a connotative linguistics which treats generation of meanings within a conceptualization not only as an extensional set of representative symbols but also as an intentional one-to-many relations to absent oppositions. That is, if an agent can identify a symbol then it is distinguishable, but the inverse is not necessarily true. Furthermore we do not presuppose a particular structure for semiotic relations. Maybe there is an evident hierarchy among symbols in terms lattices or layers but we claim that such hierarchies, if they exist, emerge as a consequence of the allocation of symbols within the structure of symbolic space that we formalize in this section. We elaborate the relationship between a negation of a symbol and its oppositions in Section 5. Until then we assume that negations do not belong to the set of symbols.

Formally, we assume that there is a symbol set $S := \{s_1,...,s_p\}$ that consists of a finite number of symbols. An opposition, O, is a binary relation defined over $S \times S$ and satisfies,

1. For all $s_m \in S$, $(s_m, s_m) \notin O$. That is the relation is irreflexive.
2. For any $s_m, s_n \in S$, if $(s_m, s_n) \in O$ then $(s_n, s_m) \in O$. That is the relation is symmetric.

First condition states that a symbol can not oppose itself since opposition relation is merely instrumental to guarantee the existence of that symbol among the others. Therefore the relation must be irreflexive. Second, we argue that if one can distinguish a symbol in the existence of the other, then the latter must be distinguishable in the existence of the former. This is a consistency principle which assumes the existence of cognitive abilities that goes beyond simple book-keeping of one way opposition relationships. Whenever a symbol enters the system, opposing to some particular symbols, system responds this perturbation as a whole and locates the incoming symbol to a corresponding setting. This reflex requires symmetricity. It also explains how the system handles granularity. That is, if a symbol is distinguishable within the existence of a set of symbols but within no proper subset of it, due to symmetricity, this set of symbols enters the system as a symbol of its own (neither gin nor tonic, a gin-tonic) and does not necessarily preserve the oppositions of its constituents.

These two properties delineate the structure of the symbol space in our model. Symmetricity condition allows us to represent the binary opposition relations as bi-directional networks while irreflexivity makes sure that there are no self

loops. Figure 1 illustrates an example of a network of oppositions for six symbols.

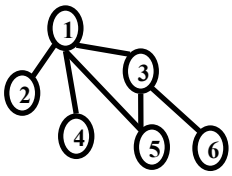

Figure 1. Example of a network of oppositions.

Apart from their existence we do not specify any valuation that infers an ordering relation over the oppositions. We claim that symbols are connected to some other symbols in such a way that their existence and identities – unified under the term "meanings" are revealed through this particular structure. Therefore meanings of a symbol comes from the oppositions of the symbol to other symbols in the same system. The most basic two premises of this statement is first, a meaning is merely a set of symbols and second, there are no opposing symbols in that meaning set.

In order to formalize, we first define opposition class of a symbol, s_n as the set $\overline{s}_n := \{s_m \mid (s_m, s_n) \in O\}$. Thus a meaning, m, is a subset of S such that if $s_n \in m$, then $\overline{s}_n \cap m = \varnothing$.

In the example illustrated in Figure 1, the set {2,4,5} constitute a meaning since it contains no binary opposing symbols. Similarly, sets {2,6}, {2,4,6} and {4,5,6} are all meaning sets among many others that can be revealed from the network of oppositions in Figure 1. However the formal definition of a meaning does not specify the symbol it explains because it is purely symmetric in terms of its constituents. In this regard we assume that individuals can generate meanings and that whenever a meaning is generated it is actually assigned to all constituting symbols.

Next we demonstrate the consequences of this assumption with an example. In this example we assume two primitive humans, a man and a woman living in the same tribe and interacting with nature. They share the same network of oppositions but they have generated different meanings as displayed in Figure 2. When these agents observe *rain* (Symbol 2) they interpret it according to their corresponding meanings. For instance man uses a single meaning, {2,3,4} and makes sense of the *rain* within the context defined by *dark* and *cold* (Symbols 3 and 4). As long as nature prove otherwise agent uses this meaning as a personal theory to understand or to make sense of *rain*. On the other hand, for the woman the situation is different because she has two meanings to interpret *rain*: *rain* means *cold* and *moon* (maybe *sun* behind the clouds looked like *moon*), on the other hand, to her *rain* also means *dark* ({2,4,6} and {2,3} respectively). This accounts for intentionality that we have discussed previously. She needs to interpret the observed symbol therefore she must get rid of or at least decrease the complexity by devising an ordering mechanism for the corresponding meanings. This ordering is contingent to her practices with the nature. In other words as the agent makes sense of the situations of observing *rain* repetitively, depending on the rewards she receives, a relative ordering over meanings is constructed and complexity is gradually reduced. However, inverse scenario is equally probable. If the collection of meanings for *rain* is incapable of making sense then the agent infers new meanings from her network of oppositions hence increases complexity. The dialectics between the urge to make sense of situations and the urge to reduce uncertainty shapes and reshapes meanings assigned to symbols. Yet we argue that the network of oppositions for an agent perturb only rarely [37]. But the system is almost always rich enough to generate new theories, new meanings to make sense of nature. In this example all symbols are observables, but in our formalization anything that can be perceived or conceived is a symbol, such as feeling of danger, love or pain and even abstract concepts such as infinity or truth.

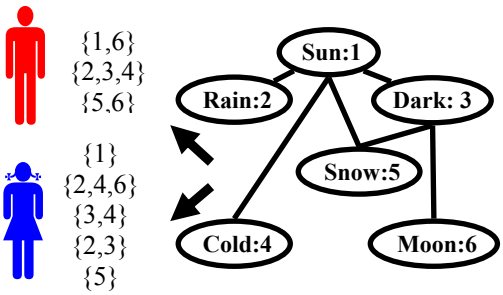

Figure 2. Two individuals sharing the same opposition networks.

Since a meaning can be conceived as a list of symbols to which it is associated, the collection of all meanings that an agents reveals from his/her network of oppositions efficiently describes all symbol-meaning correspondences. Furthermore if this collection covers the symbol space completely then it is the personal theories of an agent about everything, her world view. We label this collection as a meta-language and formalize it.

We start by the usual definition of a cover of a set. A collection of sets $X = \{x_1, \ldots, x_g\}$, where each $x_i \subseteq S$ is said to be a cover of S, if for all elements $s \in S$, there exist at least one $x_i \in X$ such that $s \in x_i$. A meta-language $M = \{m_1, \ldots m_k\}$ is a set of meanings for the network of oppositions (S, O) that covers S.

Meaning generation is an irreducible operation since it is crucial to conform to the network of oppositions as a whole. As the number of symbols get large it becomes harder to check for the consistency argument that no opposing symbols belongs to the set. On the other hand since complexity is not related with the cardinality of meaning sets but their many to one assignments to a symbol, large meaning sets comes into existence as immediate consequences of the dialectics between sense making and complexity reduction with the trade off of increased difficulty in construction of the meaning set. The situation gets worse for the generation of a meta-language that consists of many such meanings. We now state and prove a theorem that explains how minimal amount of cognitive capacity would be enough to generate a meta-language in a huge symbolic space such as the social space. The principle idea is developing base sets by which all meta-languages can be constructed, more or less in the same manner one can construct any vector in a vector space by using unit vectors. The base sets

used to construct meta-languages is called meta-language generating set.

Given an opposition relation O, over a set of symbols S, a meta-language generating set, $\mathfrak{M} = \{\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_k\}$ consists of subsets $\boldsymbol{m}_i \subseteq S$ such that for any meta-language M of (S, O), and for any $m_i \in M$, there exist at least one $\boldsymbol{m} \in \mathfrak{M}$ such that $m_i \subseteq \boldsymbol{m}$ and for any $p \neq q$, $\boldsymbol{m}_p$ is not a proper subset of $\boldsymbol{m}_q$.

Like all meta-languages the elements of meta-language generating set $\mathfrak{M}$ are sets of symbols. Also if M is an arbitrary meta-language and for any $m_i \in M$, m_i cannot contain any $\boldsymbol{m}_i \in \mathfrak{M}$. The following theorem not only guarantees the existence of a meta-language generating set for any symbol set and binary opposition relation pair but also proves that this set is unique.

Theorem : Any opposition relation O over a set of symbols S defines a unique meta-language generating set $\mathfrak{M}$.

Proof: Proof constructs set $\mathfrak{M}$ and shows that it is unique. Let $\bar{O}$ define the complement of the binary opposition relation O, $\bar{O} = \{(s_i, s_j) \mid (s_i, s_j) \notin O\}$. Let $\mathfrak{G}$ be the undirected graph representation of $\bar{O}$.

Proof continues in graph theory framework. Construct $\mathfrak{M}$ such that $\boldsymbol{m} \in \mathfrak{M}$ if and only if $\boldsymbol{m}$ defines a set of nodes that forms a clique in $\mathfrak{G}$. A clique C is a subgraph such that each node is connected to every other node and the set is maximal with respect to this property. Clearly, there may be more than one clique within a graph.

Thus each element of $\mathfrak{M}$ contains cliques as sets of nodes. Since the set of cliques of a graph is unique and is a cover, $\mathfrak{M}$ is unique and is a cover of S. From the construction of $\mathfrak{G}$, $\mathfrak{M}$ is a meta-language.

It only remains to show that if M is any meta - language of (S, O), and $m_i \in M$ then $m_i \subseteq \boldsymbol{m}$ for some $\boldsymbol{m} \in \mathfrak{M}$. In other words any meta-language contains sets that are subsets of some elements of $\mathfrak{M}$. Since there can be no opposing pairs within m_i, it follows that a non-opposing graph representation of m_i is a complete sub-graph of $\mathfrak{G}$. If m_i is not contained properly in another complete sub-graph then m_i is a clique so, $m_i \in \mathfrak{M}$. On the other hand if m_i is not a clique then it must be contained in some clique $\boldsymbol{m} \in \mathfrak{M}$ such that $m_i \subseteq \boldsymbol{m}$. This completes the proof. On the computability side, Bron and Kerbosh [3] developed an efficient algorithm to extract cliques of a graph and the issue is still live in computational graph theory.

According to the theorem, meta-language generation runs over complete sub-graphs of non-opposing symbols. In other words, in a world populated with lots of symbols, locally non-opposing symbol domains would be sufficient to generate meanings and hence meta-languages.

In the perspective we put forward here, nature is a huge system of symbols where scientific disciplines like Physics or Biology and many others search for existences, theories or regularities which we insist on unifying under a single term meanings. With this stance, we also argue that meanings are generated in order to identify features that capture important characteristics in an efficient way. This connects meta-language model to Sapir-Whorf hypothesis that distinguishes language as the basis of interaction. Goldspink and Kay [19] state that language has significant implications for the dimensionality of the resulting higher order structures it can generate and support. For instance whenever we notice that a collection of water molecules flow as a "fluid", we simplify its description considerably [12]. Consequently, language brings its own symbols into our symbolic space redefining granularity and corresponds to meanings that makes description simpler. Thus language is not simply a renaming of symbols but it is more.

The entire thrust of the formalization is to ground this perspective to make it implementable in a working model of sociality. Since a realistic model would consist of agents that are heterogeneous in their meta-languages, social simulation methodology fits best to our purposes. In the next section we propose a simulation based implementation of meta-language where agents do not interact merely with their immediate neighbours but with all populated agents randomly – offset with a function of their distance.

3 IMPLEMENTATION OF META-LANGUAGE

We have already presented an implementation of the meta-language model for the case of interactions with nature. Social interactions, on the other hand, are considerably different since they involve two reflexive parties and reflexivity happens in and through language[20].

Obviously social interactions can not be reduced to pure symbolic exchanges of curiosity and complexity reduction. Social interactions are also structured by other aspects of the environment such as resource distribution, spatiality and power relations. Thus semiotic structure is different from economic, geographic and political structures which also inform social interaction. However, even if an action is determined to a large extent by some sources, these sources would still have to be decoded into a meaning in meta-language.

Meta-language allows agents to play on the multiple meanings of symbols, or in other words stir their imagination, in such a way that agents may redefine situations in ways that they believe will favour their intents. Situations, evaluation criteria and intent of each agent, on the other hand are defined by the context of a completely foreign dynamics, such as economics or politics.

As agents develop their own meta–language models, social interactions loads inherited information to code. Hence for any agent, symbol-meaning correspondence is dynamical in nature. A symbol might not only correspond to multiple meanings but also an agent might attach a new meaning to a particular symbol or drop a meaning from it. This requires the existence of a reliability measure for meanings that inflects or deflects according to rewards in social interactions which we briefly note as the mechanism governing the protocols of inherited information. This cognitive mechanism bridges social space to meta-language.

Should such a mechanism be symmetric for benefits and losses? Should it depend on absolute magnitudes or should it depend on relative increments or decrements? This mechanism

models a cognitive process and as DiMaggio [13] states, it must be consistent with results of empirical research on cognition.

Kahneman and Tversky [24] argued that individuals are tuned to relative changes rather than absolute magnitudes and that valuation is not symmetric for decrements or increments. They supported this view by numerous cognitive experiments. In this vein they have developed Prospect Theory of decision making. Since reliability is in essence a valuation procedure over meanings, adopting Value Function calculus from Prospect theory as an updating mechanism fulfils the requirements that DiMaggio [13] emphasize. Such a mechanism is of the form,

$$r_{t+1} = \begin{cases} r_t + e^{\alpha} & if\ e > 0 \\ r_t - \lambda(e)^{\beta} & \text{otherwise} \end{cases}$$

where, r_t and r_{t+1} are a reliability for some meaning, at time t and t+1 respectively and e is a function of increments and decrements that realize in some other sphere. Default characteristic of value function dictates that decreases are steeper than increases.

Hence, mechanism that bridges social space to meta-language can be implemented by a Value Function calculus adopted from Prospect Theory. Yet, this not the only way; other experimentally and empirically derived models of action embedded within different contexts are equally valid. The inverse direction tells us how meta-language informs social interaction.

Durkheims' structure of beliefs versus forms of social organization, Marx' ideology versus social formations, Weber's cultural forms versus power relations and modes of economic organization and many others show that much of the history of social theory has been organized around the debate of the connection between symbolic systems and other levels of social life. However our objective in this article is to develop a model that informs us how actors define social situations in which they find themselves, but not the actual practices these definitions would imply.

In its most basic form, a social interaction is initiated when an active agent selects a target agent who plays a passive role. These agents can be arbitrarily heterogeneous both in their opposition networks and meta-languages. With the intentions defined by the environment which he can only make sense through his meta-language, active agent signals its intents with a symbol and a context within which the passive agent is expected to make sense of the signalled symbol. In this way, social interaction is defined as capability of mutual sense making.

We present two mechanisms that implement this perspective. When passive agent receives a signal from the active agent in the form of a symbol within a context or in brief as a semiotic code, he tries to decode it referring to his meta-language. In the first mechanism, only mutual existence of the same meaning (context) for the signalled symbol is necessary and sufficient for the interaction to occur (Figure 3a). So, in order to interact, passive agent must already be equipped with the intention that triggered active agent. This is in line with the intentional arc principle assumed by Dreyfus. "The intentional arc names the tight connection between body and world, such that, as the active body acquires skills, those skills are 'stored', not as representations in the mind, but as dispositions to respond to the solicitations of situations in the world." [14; p.362]. Therefore, active and passive identities are purely inconsequential. Such a mechanism mimics a tag model with zero tolerance [23,22], where the tag is not constant but is actually coming from a system of symbols which has its own dynamics. In return two agents can interact over more than one symbol. This property enables the researcher to control not only the strength of ties between agents but also asymmetric relationships in the simulation that we elaborate in the next section.

Second mechanism relaxes mutual existence constraint and loads passive agent a cognitive capacity. When passive agent receives the semiotic code, instead of searching for the exact context for the symbol, he tries to infer it from the available contexts in his repertoire. This could be achieved either by particularization or generalization of meanings previously attributed to the symbol (Figure 3b). This corresponds to maximal grip, "the body's tendency to refine its responses so as to bring the current situation closer to an optimal gestalt." [14; p.362]. Both of these operations respect the network of oppositions of the passive agent. Therefore with this mechanism sense-making is defined as recognizing the same set of oppositions for a specific symbol. Cognitive capacity of the passive agent can also be controlled by the researcher depending on the target of the simulation.

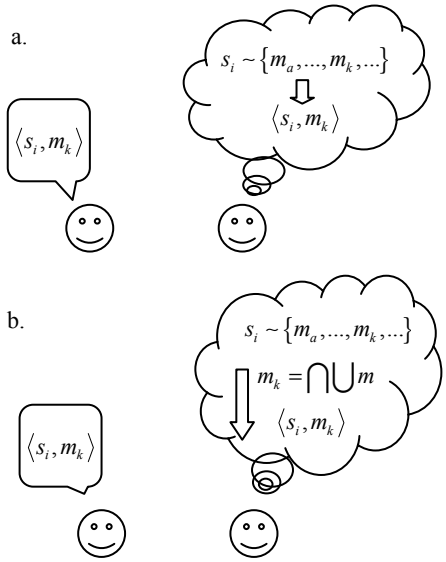

Figure 3. Two example mechanisms of sense-making through meta-languages.

However, semiotic competence does not imply that agents agree in their moral or emotional evaluations of given symbols. While sense making is governed by meta-language, evaluations of symbols is an outcome of the social interaction process and belongs to the practices in social space.

The mechanisms governing the evaluations of symbols bridges meta-language to social space and can be implemented in a plethora of ways. For instance, given that a social interaction is initiated, social practice might follow as an independent procedure which in return allows the researcher to study sense making as a constraint on the dynamics of target emergent phenomena such as cooperation. This idea mimics Gilbert's [18] "interaction" extension of Schelling's [32] model. There the tags that mark the differences between agents vary thus as the simulation proceeds the agents themselves decide which of all their tags become their significant characteristics. It could be

"ethnicity" or "gender" or something else. Similarly in our model, as agents interact with each other where they evaluate symbols through a game whose rules are set in an independent domain, cooperating local semantic societies are expected to emerge. This claim has strong supports from adaptive coin flipping theory [10,17]. In this case meta-language model extends social simulation methodology by providing the researcher a toolkit to analyze the factors of heterogeneous network formations and their relative stabilities. Same idea goes for the inside of the emergent social networks as well. Social networks are neither totally closed [8] nor random [39]. Agents allocate positions within a network that provides unique opportunities and access to these opportunities [38]. Then, how can some agents allocate superior positions than others such as structural holes [5]? Does this mean that agents within an emergent social network are still heterogeneous enough to generate some specific network structures? What are the dynamics of non-complete but still embedded social networks? We claim that social simulation models that implement meta-language will provide powerful insight for such questions.

A second mechanism that bridges meta-language to social space considers a priori knowledge of the researcher about the target population. Simulation methodology is also very powerful in studying "what-if" scenarios. In situations where a causal relationship between a symbol within a particular context and a specific behaviour pattern is obtained [40], it is possible to associate them in the model. In this regard, meta-language model complements the intervention model developed by Kay et al. [25] which predicts effectiveness of an intervention within an organizational domain in terms of the interplay between individual agency and institutional structure.

As a last note for this section, although we have argued that mechanisms of meta-language model can be implemented like a tag model, conceptually they are completely different and resemblance is only in technical accounts. More specifically, while in tag models agents are equipped with a priori characteristics, in meta-language models agents are equipped with allocations for the flow of knowledge that bridges meta-language and social space.

4 EMERGENCE

If emergence is a function of interactions and/or structural coupling of some resulting networks [30], then this structural characteristic must have been stored in the meta–languages of the constituting agents. Otherwise their meta–languages block interactions and prohibit the "desired" structure to emerge. With the first bits of the emergent phenomenon coded in meta-languages, convergence accelerates due to snow ball effect of social causation. In return heterogeneous groups of relatively homogenous constituents emerge. In a social simulation model where all agents can interact each other rather than their immediate neighbours, between group interactions would best tell the effects of social causation.

Adaptive coin flipping theory tells us that in a meta-language based artificial society, sub-networks with strong ties would emerge. Within such a sub-network, where agents interact with each other over multiple symbols, some particular meanings suppress others through repeated interactions. In this case, constituting agents make sense of some symbols over unanimously agreed contexts. Thus complexity is reduced and a common ideology is emerged. In return, this common ideology constraints the constituting individuals in such a way that, membership to the society would be contingent to complying to the ideology. Yet, no two individuals can be assumed to be identical in their network of oppositions in the existence of tremendous amount of symbols. Therefore the group ideology never covers the whole set of meta-language. Ethos remains as the seeds of the prospective emergent structures. Meta-language models capture these features and provide the researcher the ability to define intentionality on instable domains.

Last but not the least, in a meta-language model, emergent social networks and ideologies around symbols come into existence simultaneously. Thus monitoring symbol-meaning correspondences serves as a detection mechanism of emergent structure as discussed in [12].

5 META-LANGUAGE IN PRACTICE

A close examination of the algebraic structure of meta-language reveals that in an attempt to generate an artificial society numerous agents can be populated from a single binary opposition over symbol sets of unrealistically low cardinalities. However more interesting applications of meta-language involves opposition networks obtained from empirical data in the form of narratives. Meta-language model acknowledges the human tendency to think in terms of oppositions. But in the model, we have ruled out the negations of symbols whereas it is clear that the exact opposite of "rich" is actually not "poor", but "not rich". Stated otherwise, a perfect dichotomy exists only between a symbol and its negation. Meta-language is not defined over dichotomies (and can not be), it is defined over positive symbols. Consequently, in a grounded symbol system such as language, empirically observable as narratives, any symbol, such as "poor" opposes symbol "rich" if and only if it is proximate to "not rich". This perspective defines the dual form of opposition relations in terms of proximities to negations and opens a gateway to make use of cognitive anthropology methodology to obtain opposition networks from narratives.

Following Sapir-Whorf hypothesis, cognitive anthropologists argue that narrative mode of thought constitutes the core of meaning generation. Clearly, individuals carry their past experiences into every decision making situation as structured knowledge in their brains. This knowledge, organized in terms of categories, influence how the individual understands the world [34]. Decision making process, especially in complex situations is a result of attaching a meaning to the symbols around the individual. Therefore, models developed in cognitive anthropology seek to reveal the way in which decision makers are making sense of the situation. Their aim is to present the structure of the decision making problem from the lens of the decision maker [15].

Specifically it is assumed that the proximity of symbols and of their meanings increases as they co-occur in an utterance [11]. This relationship has been used to develop taxonomies, clusters or causal relationships that explain how individuals attach meaning to symbols. Since symbols are elaborated at an higher level through pooling, relationships at symbolic level are lost due to aggregation methods such as principle component analysis. Furthermore all of these techniques require the subjective interpretation of the researcher such as inferring the

tone of the narrative data [6] and cannot be used to generate opposition networks.

On the other hand when narrative data is coded with the original tone it represents, unavoidably symbols and their negations are recorded as distinct codes. In this way a narrative data such as “He is not rich. He is a scientist.” increases the proximity of “scientist” with “not rich” but does not effect its proximity with “rich”. Then using the dual definition of opposition and a measure such as Jaccard’s coefficient [41], proximity of “scientist” and “not rich” reveals the opposition between “scientist” and “rich”. In that way it is possible to construct opposition networks of individuals from narratives and study their dynamics under domains of interests or with what-if scenarios within a social simulation model.

6 CONCLUSIONS

The dialectic of emergence and social causation remains one of the major issues in social science and we agree with Castelfranchi [7] that social simulation models offer helpful perspectives in this debate. However the principle methodological problem for social simulation models is the representation of knowledge about the structure. Social action depends on the knowledge about all levels of emergent structure and this knowledge is not limited to but produced by the local binary interactions of the individual agents. This knowledge resides in the impersonal area of symbol systems and is produced by a meta-language which carries the knowledge of emergent structure beyond local networks to the topological knowledge of the macro phenomena. Individual agents use these symbol systems as tool-kits to make sense of symbols, and use their world-views to ground the symbols and ethos to challenge meaning-symbol correspondences through their everyday practices. Symbolic interaction is not only the symbolic affirmation of shared social classifications and normative protocols that regulate interaction but is also making sense of expressive, symbolic behaviour and decoding the intent of the counterparts from these symbols. Meta-language is the key concept to understand how this knowledge is generated through semiotic relations. Albeit the theoretical support from linguistics and culture studies, a compatible proper formalization for a meta-language model through which agents can encode and decode knowledge about the structure has never been made. This paper develops the formal infrastructure of such a model and elaborates various mechanisms that can be implemented within a social simulation model.

The definition of opposition between symbols has also a dual form as proximity to negations. This form is particularly useful in deriving opposition networks from narratives that can be used for dynamic analysis and to test what-if scenarios within a social simulation model. In that way, meta-language model developed in this article promises opportunities to social simulation modellers to bridge the gap between what is aggregate and emergent and what is individual specific. We have also discussed that meta-language models are useful in explaining the heterogeneity eminent in social networks and provides bases for the emergence and dynamics of embeddedness. Our methodology also has strong implications for validation of simulation models as well. But due to space constraints we leave this subject for further publications.

REFERENCES

[1] Anderson, M. L. Embodied Cognition:A field guide. *Artificial Intelligence*, 149 2003, 91-130.

[2] Beer, R. D. Autopoiesis and Cognition in the Game of Life. *Artificial Life :* , 10 2004, 309-326.

[3] Bron, C. and Kerbosch, J. Finding all Cliques of an Undirected Graph. *Communication of the ACM*, 16 1973, 575 - 577.

[4] Brooks, R. Intelligence without representation *Artificial Intelligence*, 47 1991, 139-159.

[5] Burt, R. S. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, 1992.

[6] Carley, K. and Palmquist, P. Extracting, Representing, and Analyzing Mental Models. *Social Forces*, 70 1992, 601-636.

[7] Castelfranchi, C. *Through the Minds of the Agents*. City, 1998.

[8] Coleman, J. S. Social Capital in the Creation of Human Capital. *American Journal of Sociology*, 94, Sociological Analysis of Economic Institutions 1988, 95-120.

[9] Conte, R., Edmonds, B., Moss, S. and Sawyer, R. K. Sociology and social theory in agent based social simulation: A symposium. *Computational & Mathematical Organization Theory*, 7 2001, 183-205.

[10] Cooper, W. S. and Kaplan, R. H. Adaptive "coin-flipping": a decision-theoretic examination of natural selection for random individual variation. *Journal of Theoretical Biology*, 94, 1 1982, 135-151.

[11] D'Andrade, R. *The Development of Cognitive Anthropology*. Cambridge University Press, New York, 1995.

[12] Deguet, J., Demazeau, Y. and Magnin, L. Elements about the Emergence Issue: A Survey of Emergence Definitions. *Complexus*, 3 2006, 24-31.

[13] DiMaggio, P. Culture and Cognition. *Annual Review of Sociology*, 23 1997, 264-287.

[14] Dreyfus, H. Intelligence without representation – Merleau-Ponty’s critique of mental representation: The relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences* . 1 2002, 367-383.

[15] Eden, C. Cognitive Mapping and Problem Structuring for System Dynamics Model Building. *System Dynamics Review*, 10, 2-3 1994), 257-276.

[16] Edmonds, B. Pragmatic holism (or Pragmatic reductionism). *Foundations of Science*, 4 1999, 57-82.

[17] Gigerenzer, G. *Rationality:Why social context matters*. Cambridge University Press, City, 1996.

[18] Gilbert, N. Varieties of Emergence. In *Proceedings of the Social Agents: Ecology, Exchange, and Evolution Conference* (Chicago, 2002)

[19] Goldspink, C. and Kay, R. Bridging the micro-macro divide: A new basis for social sciences. *Human Relations*, 57 2004, 597-618.

[20] Goldspink, C. and Kay, R. Systems, Structure and Agnecy: A Contribution to the Theory of Social Emergence and Methods for Its Study. In *Proceedings of the ANZSYS Conference-Systemic Development:Local Solutions in a Global Environment* (Auckland, New Zeland, 2007)

[21] Habermas, J. *The Theory of Communicative Action*. Polity, City, 1984.

[22] Hales, D. and Edmonds, B. *Evolving Social Rationality for MAS using "Tags"*. ACM Press, City, 2003.

[23] Holland, J. *The Effect of Labels (Tags) on Social Interactions*. City, 1993.
[24] Kahneman, D. and Tversky, A. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, 2 1979, 263-292.
[25] Kay, R., Goldspink, C. and Preston, A. Organizational Change: Revealing the micro-macro patterns underlying social system dynamics in a financial services context. In *Proceedings of the 1st ICC Workshop on Complexity in Social Systems* (Lisbon, 2008)
[26] Kravchenko, A. V. Toward a Bio-Cognitive Philosophy of Language. *Journal for Interdisciplinary Work in the Humanities* 2002.
[27] Kravchenko, A. V. Essential properties of language, or, why language is not a code. *Language Sciences* 29 2007, 650-671.
[28] Luhmann, N. *Social Systems*. Stanford University Press, Stanford, CA, 1995.
[29] Merleau-Ponty, M. *Phenomenology of perception* Routledge, London, 1996.
[30] Sawyer, K. The Mechanisms of Emergence. *Philosophy of the Social Sciences*, 34 2004, 260-282.
[31] Sawyer, R. K. Artificial societies: Multiagent systems and the micro-macro link in artificial society theory. *Sociological Methods & Research*, 31 2003, 325-363.
[32] Schelling, T. C. Dynamic Models of Segregation. *Journal of Mathematical Sociology*, 1 1971, 143-186.
[33] Schutz, A. and Luckmann, T. *The Structures of the Life-World.* Northwestern University Press, City, 1973.
[34] Senge, P. M. *The Fifth Discipline : The Art and Practice of Learning Organization*. Currency Doubledaly, New York, 1990.
[35] Silverman, H. J. *Derrida and Deconstruction* Routledge, New York 1989.
[36] Sonesson, G. *Opposition*. City, 2004.
[37] Swidler, A. Culture in action: Symbols and strategies. *American Sociological Review*, 51, 2 1986, 273 - 286.
[38] Uzzi, B. The Sources and Consequences of Embeddedness for the Economic Performance of Organizations: The network effect. *American Sociological Review*, 61 1996, 674-698.
[39] Watts, D. J. Networks, Dynamics, and the Small-World Phenomenon. *American Journal of Sociology*, 105, 2 1999, 493-527.
[40] Wellman, M. P. Inference in Cognitive Maps. *Mathematics and Computers in Simulation*, 36 1994, 137-148.
[41] QDAMiner-Provalis Research, `http://www.provalisresearch.com/QDAMiner/QDAMinerDesc.html`

A Brief Survey of Some Results on Mechanisms and Emergent Outcomes

Bruce Edmonds
Centre for Policy Modelling
Manchester Metropolitan University

Abstract. The mechanisms/abilities of agents compared to the emergent outcomes in three different scenarios from my past work is summarised: the El Farol Game; an Artificial Stock Market; and the Iterated Prisoner's Dilemma. Within each of these, the presence or absence of some different agent abilities was examined, the results being summarised here – some turning out to be necessary, some not. The ability in terms of the recognition of other agents, either by characteristics or by name is a recurring theme. Some of the difficulties in reaching a systematic understanding of these connections is briefly discussed.

1 INTRODUCTION

Put briefly the question of this workshop is: *what agent abilities are necessary for which emergent outcomes*. The importance of this can be seen in two ways: *firstly*, which of our abilities have evolved due to the advantage they give us in terms of social organisation (the *social intelligence hypothesis* [26]); and *secondly*, given the abilities we have, what kinds of social structure *could* emerge – i.e. what is the social *scope* of our species. To get a handle on this we need to understand some of the connections between agent abilities and what this means in terms of possible outcomes. Since this involves questions of scope and counterfactuals, a look at the evidence of one instance of a social species at one point in time is clearly not sufficient. Social simulation models provide one way of exploring this (examining a range of species and societies and gathering evidence is another).

Since this is a question I have thought about and investigated over at least 10 years, in a number of simulation models, I summarise some of these results here in brief. These models include both positive and negative results – that is, simulations in which certain abilities do seem to be necessary for certain outcomes and those in which they turn out not to be necessary. The paper is structured into the different games/environments in which the agents are placed. It must be noted that all these are very much simpler than those of a "complete" society and thus can be only taken as indicative of what might be the case there.

These are grouped together in terms of the kind of interaction the agents are put inside, with subsections for each mechanism investigated. I end with a short summary and a brief discussion of the difficulties of trying to work out what abilities cause which emergent outcomes and why.

2 EL FAROL BAR (OR *MINORITY*) GAME

In the El Farol Bar game [1] there are a fixed number of individuals (usually 100) have to decide each week whether they go to the El Farol Bar, or not. They want to go if it is not too crowded (more than 60) go but not otherwise. Each individual knows the past history of the attendance there and has to predict this week's attendance and decide accordingly. The result is that the attendance oscillates wildly (within a broad range of parameter settings) around the critical point. This was re-christened as the "minority game" in [3].

In [8] and [7] I extended this model to allow: sophisticated communication; a social network; open-ended learning; and strategies that allowed individuals to take their cue from specific others rather than just base their guesses on the total attendance numbers (as in the above version). This allows individuals to take their cue from identifiable others (I will go to the bar this week if agent-23 did last week, etc.).

2.1 Sophistication of Prediction Strategies

What Brian Arthur (and others who have investigated this model) found is that it does not matter which prediction strategies are available to the individuals [1] – the same outcomes seem to result as long as there is sufficient variety of strategies among the population of players]. Thus it does not matter whether the strategies for predicting patters in the attendance are sophisticate or simple, the structure of the game means that successful strategies are self-defeating after a short while. However this does seem to rely on a rough parity between cognitive power between the participants.

2.2 Open-endedness of learning ability

I the presence of certain kinds of open-ended learning, where strategies can be arbitrarily elaborated and individuals have a limited ability to search for new strategies this can lead to a spontaneous emergence of heterogeneity in terms of the kinds and styles of strategy developed by each participant. How different theses strategy styles are depends on the range of strategy primitives available to an individual (the language that strategies belong to).

2.3 Ability to recognise individuals

The ability to explicitly recognise other individuals (i.e. by "name") seemed to have several effects. I did allow the population as a whole to develop a better population of

strategies, in that it seemed able to facilitate the *dis*coordination of decisions to attend between agents. At the same time it seemed to produce a greater variety of strategy fitnesses within agents [8]. Thus, as part of the expressiveness of the strategy space available to individual's this also had a similar effect to that described in section 2.2 above.

More importantly this greatly increased the amount of social embedding (in the sense of [19]) that seemed to occur – that is the extent to which different agent's strategies become interdependent on the results of others' strategies etc. Such embedding has positives and negatives from the point of view of the individual – it means that information is efficiently used within the "society" of individuals, so that newcomers/outsiders are usually at a disadvantage since they have not learnt to exploit this resource. However occasionally such outsides might do much better than any because they can more easily discover totally new strategies and others will take a while to learn how to use its decisions [7].

2.4 Imitation

The ability to imitate a strategy is different from observing someone else's behaviour and using it as an input to one's own decision making strategy. It is an ability to copy the strategy behind the observed behaviour and thus reproduce that behaviour. In this scenario, imitation did not seem to facilitate either the general (dis)coordination of behaviour or the social embedding. Maybe this is due to the nature of the task which favours anti-coordination and a dissimilarity of strategies between individuals.

2.5 Communication

Similarly to imitation, described above, communication between agents in this model does not seem to have a crucial impact on the outcomes in this model. Since communicating in this model gives another individual additional information that it could use to its advantage (and the communicating agent's disadvantage) *if* it does contain a message that can be correlated with the sender's actions then it is so the advantage of the sender to "fool" the receiver, by constantly changing that message. Thus in some of the simulation runs agents were developing utterances of the form "*not not not not ... not not what I did last time*" as they appended "not"s in an effort to reverse the meaning of their communication. In summary the situation here is not conducive to the emergence of communication since it is to an agent's disadvantage to say anything that is meaningful in terms of its own actions.

3 ARTIFICIAL STOCK MARKET

In this situation there are a number of traders and a single market maker that sets prices and buys and sells the various stocks. Thus each trader can hold different mixes of cash the the various stocks available. Each trader has a small regular external income and aims to make money by investing or speculating. Each stock has a fundamental in terms of an dividend rate which follows a slow random walk – the trader gains interest from holding the stock in terms of this fundamental. There is a trading cost, so that it is not a good idea to be over-trading. The market maker is under obligation to buy stocks from traders at the current price or sell what stock it has, but it also sets the prices. The prices are set according to a simple rule which compares supply and demand – if supply is greater than demand (more want to sell than buy) the price goes down, and in the opposite case upwards. This is basically an extension of [27]

This situation is similar to that of the El Farol bar game in that traders are competing to "out-guess" each other. It is a good idea to start buying a stock after and when others have been selling them and to sell after others have been buying into it. If one just follows others, one step behind, buying when others have bought and selling when others have sold last time you just lose money. If the market prices are fairly stable, it is best to buy and hold the stock with the highest dividend rate, but if dividends are low and prices volatile, it might be possible to make more money by speculating. Thus this is a fundamentally competitive situation, and chances for cooperation are minimal.

3.1 Type of Learning Mechanism

In I investigated what happens if I swapped two different learning algorithms, but ones with the same inputs, outputs and rough expressive power. One was a neural network and one a genetic-programming algorithm. They both had a similar theoretical ability in terms of the functions between inputs and outputs that they could learn. However they did have different characteristics. The NN learns more smoothly, sampling intermediate functions and values from one strategy to another. The GP algorithm encodes rougher approximations and is able to change more sharply from one strategy to another in its memory.

Since a stock market model like this one acts to "amplify" and "react" to sudden changes running the model with each of these as learning mechanisms makes a lot of difference to the outcomes. The version with NN learning reacted smoothly and gradually, showing gradual adaption of prices and long-term price "waves", whilst the GP version showed sudden, clustered volatility and characteristic speculative bubbles. Thus we can conclude (unlike others) that in some situations the exact cognitive model can make a crucial difference, even if they have the same "abilities".

3.2 Anticipation

There are two different kinds of feedback that can be observed when one has taken an action or implemented a strategy – the success of that strategy (which explicitly or implicitly requires comparison with a goal and may be given a numerical measure such as a *fitness*) and the how well it results in the outcomes that were predicted (the *accuracy*). This corresponds to *instrumentalist* and *realist* strategies of learning, respectively: the former simply using raw feedback according to success to learn what to do; and the later which learns to make good predictions about the effects of actions (or causes) using feedback as to their accuracy and then uses these to reason about which action or strategy should be chosen assuming these that will best achieve their goal. It is generally supposed that the later kind, labelled anticipatory, will be more sophisticated but require more computational resources (although there are leaner versions of this, e.g. [28]). It is clear that sometimes such anticipation is needed, but that at other times it is not.

Thus I mixed traders which uses an instrumental style of trading (the fitness of just what worked recently and the content just being the immediate buy/sell decision) with one which evaluated GP expressions as to how accurately they predicted

prices (recently), the currently best being used to predict the next price changes and hence determine its buy/sell decisions. What I found in this model [10] is that anticipation did not help a trader earn more money in the long term, but have a different pattern of trading. The anticipatory traders would detect patterns in the short term and for a while do better than the instrumental traders, however the market would then change unexpectantly due to the instrumental traders discovering better strategies (due to the drain on them extracted by the anticipatory traders) and the anticipatory traders would then lose a lot of money, until they recognised their predictions had failed. Thus anticipation did not make them better traders in the long run.

3.3 Context Awareness

Learning in a context-aware manner means that there is a two stage process to learning: *firstly*, recognising the appropriate type of context; and *secondly*, learning assuming that context. This *does* depend upon the assumption that the situations being learned about *do* divide usefully into a series of recognisable contexts [6]. The folk theory of traders does suggest that they recognise and respond to market "moods", so it seems plausible to suggest that context recognition might be useful for a trader in such a market.

[9] suggests an algorithm for simultaneously learning the appropriate contexts and the knowledge within these contexts. This was tried in a stock market model in [11]. Agents that used context-dependent learning was compared with those that that used a similar style of learning but assumed there was only one context to learn within (effectively being context-free learning).

Preliminary results did seem to indicate that context-aware traders did better than context-free traders.

4 PRISONER'S DILLEMA GAME

Social dilemmas are when there is an outcome that is desirable for all (usually couched as when everyone *cooperates*), where individuals can do even better for themselves if they act selfishly (usually *defection*) but if everyone does so everyone does very badly. The situations characterised as the "Tragedy of the Commons" [24] are a classic example of this. Clearly if everyone just acts myopically in their self interest, then the worst outcome is inevitable.

In order for this to be avoided some additional mechanisms or social structure is necessary. Over the years quite a number of different ways of achieving this basic type of cooperation have been developed, including: kin selection, iterative play with memory of past interactions with each player, enforceable contracts, and group formation.

The version of this situation studied here is when each interaction individuals play a number of rounds of the prisoner's dilemma game, and the individuals are propagated into the next iteration with a probability that depends upon their total score at playing the game. Some of the work in this area is summarised in [18].

4.1 Tags

The ability to recognise whether the person you are playing with will be cooperative is obviously advantageous to an individual in this game. It would also be useful to be able to recognise and remember every individual so that if you met them again you would have information about their behaviour. The former of these is impossible in most situations and the later expensive in terms of cognitive resources and impractical in large populations.

However it turns out that simply being able to recognise some observable features of individuals and preferentially chose those similar to oneself can help establish cooperative groups, even where these features have no necessary link to behaviour. This was identified and called "tags" in [25]. In other words a high over-all level of cooperation is maintained even where this is not an evolutionary stable strategy. What seem to happen is the following:

1. A small group of co-operators with the same (or similar) tags happen to form;
2. Since individuals in this group preferentially play each other, they outperform those in the general population and are preferentially populated into the next iteration, so the group grows;
3. Eventually a defector appears in the group (due to an invading defector or a mutation) and does even better than the others in that group at their expense and hence multiples faster than the others;
4. The group becomes dominated by defectors and the fitness of the members rapidly decreases, leading to the death of its members since they now score less than those in the surrounding population.

Thus although cooperation in these groups only lasts in the medium term, there is continual arising of new groups to replace those that become "infected" by defectors.

4.2 Link Imitation

In a social network where a link indicates whether the nodes will play together or not, any "groups" are not indicated by similarity of tags, but rather by the structure of the links themselves. An effect similar to the one in 4.1 above can be observed where nodes compare their success with other nodes and their links and strategy copied if they are doing better than themselves. This is the SLAC algorithm of [21].

As above this can allow the network to structure itself, by breaking into groups or otherwise partially isolating defectors, and thus maintaining cooperative groups or regions of the network in a dynamic fashion.

5 OPINION DYNAMICS

Opinion dynamics refers to a wide range of models, where there are a set of individuals each of which has a particular opinion (from a given range) at any one time. Over the simulation these individuals interact in a pair wise fashion (either randomly chosen pairs or restricted by a given social network) in which the opinion of one node affects that of the other to make it more similar to its own and to have a greater impact the closer or more coherent are the two individuals' opinions. The overriding dynamic in such models is the clustering of individuals into "groups" with similar opinions – the key questions being how many clusters form, how long does it take and how stable they are. The original and best know of these is the family of models starting with [4].

Departing from the spirit of the [5] model, is a new, simpler model. This has a fixed number of nodes and directed arcs

between these. Each node has a numeric value representing the strength of its belief on a certain issue as well as a value representing its susceptibility to influence by another. Each iteration of the simulation a random arc is selected and the opinion of the node at the origin of this is copied into that of the destination with a probability of its susceptibility. Thus eventually (without noise and give the network is connected) all nodes will have the same opinion and change will cease.

An elaboration of this model is designed to model the process of consensus formation among agents. Here the opinion of each agent can be thought of as a binary string, where each bit indicates the belief (or not) of each of a sequence of possible beliefs. There is a consistency function from possible bit strings to the interval [-1, 1] that indicates the consistency of this set of beliefs. The copy process involves the copying of a single bit from origin to destination according to a probability determined by the change in consistency that would result in the destination node. There is also a "drop" process done by single nodes which may drop a belief according to a probability related to the change in consistency that would result from this. Thus it is much more likely that the consistency of the beliefs in each node will increase over time, and also that nodes with similar beliefs will be clustered together. This model is described more in [14].

5.1 Topology

One surprising and quite general result is that, for a broad range of topologies (i.e. those that naturally occur in social networks – that are connected and with a relatively small diameter) the topology does not seem to make much of a difference to the consensus formation process. This has, in fact, been proved in a slightly abstracted general model.

5.2 Belief rejection

In the above model the effect of the presence of the ability of agents to drop beliefs is obviously important, otherwise eventually all nodes will end up with all possible beliefs (depending a little on there always being a positive probability of the copy process occurring regardless of the resulting change in consistency).

6 CONCLUDING DISCUSSION

Clearly whether the abilities of agents are critical for the emergence of certain emergent outcomes depends upon the situation that the agents are within as well as the way in which they interact. Thus a general pattern of necessary abilities is not clear. However, in the above results the ability to recognise social clues and identities is a recurring theme, and is clearly necessary in some instances.

There is a general tendency to always argue that more abilities or more sophisticated abilities are necessary, on the grounds that one can almost always conceive of a situation where it would be required. This is plausible in that we know of the wide range of abilities that humans possesses, so it is natural to suppose that these are necessary.

However this need not be the case. *Firstly*, it is probably that nature evolves multiple and overlapping mechanisms to achieve any particular purpose, since this makes for a more robust and certain result in an uncertain world, so just because humans possess an ability does not mean that it is necessary (though might indicate it is helpful). *Secondly*, it assumes that human intelligence is a general ability, which requires a set of features to be obtained but is then sufficient. However, as I argue elsewhere [12], any intelligence will have different pros and cons, be better for some tasks than others – i.e. a *general* intelligence is impossible (unlike a general computation device, which clearly *is* possible [29]).

Another difficulty with this kind of exploration is that the advantage (or even necessity) of an ability may only become apparent in the presence of certain emergent social structures. Thus there may well be a processing of social boot-strapping that must occur if certain other social structures or other phenomena will be developed.

If one attempts to evolve whole societies from scratch and one is successful in this, one learns one set of abilities that are necessary, but one still does not know if there are other (e.g. simpler) ways of achieving the same result or whether *all* the abilities are indeed necessary.

The fact that some of these social structures do seem to be emergent makes the prediction of when they will result and when not almost impossible to tell without looking at case studies or doing extensive simulation exploration. Thus the problem of studying the connection between abilities and emergent outcomes is an extremely hard one.

One thing that *is* needed is the systematic recording of simulations and results, so that some of the conditions and connections can be started to be mapped. There *are* models that share some almost standard: kinds of interaction, topology of interaction, abilities of agents, temporal structure (e.g. evolutionary propagation), system goals, etc. or at least are various of these relative to a common ancestor (e.g. a PD game). A website that mapped versions of these things (e.g. different versions of a PD game, or topology) and then linked these together to a simulation that brought them together (with the various results) would start to put this jigsaw together.

7 REFERENCES

I apologise about the number of citations to my own work, however this seems unavoidable in a summary of my own results. Versions of almost all my papers are accessible from the URL: http://bruce.edmonds.name/pubs.html

[1] Arthur, B. (1994). Inductive Reasoning and Bounded Rationality. American Economic. Association Papers, 84: 406-411.

[2] Axelrod, R. (1984) The Evolution of Cooperation, Basic Books, New York.

[3] Challet, D. and Y.-C. Zhang, Emergence of Cooperation and Organization in an Evolutionary Game. Physica A, 1997. 246: p. 407. http://xxx.lanl.gov/abs/adaporg/9708006.

[4] Deffuant, G, Neau D, Amblard F, and Weisbuch G (2000) Mixing beliefs among interacting agents. Advances in Complex Systems 3:87-98.

[5] Deffuant, G., Amblard, F., Weisbuch, G. and Faure, T. (2002), How can extremism prevail? A study based on the relative agreement interaction model. Journal of Artificial Societies and Social Simulation, 5(4) http://jasss.soc.surrey.ac.uk/5/4/1.html

[6] Edmonds, B. (1999) *The Pragmatic Roots of Context*. CONTEXT'99, Trento, Italy, September 1999. Lecture Notes in Artificial Intelligence, 1688:119-132.

[7] Edmonds, B. (1999). Capturing Social Embeddedness: a Constructivist Approach. Adaptive Behavior, 7:323-348.

[8] Edmonds, B. (1999). Gossip, Sexual Recombination and the El Farol Bar: modelling the emergence of heterogeneity. Journal of

Artificial Societies and Social Simulation, 2(4). (http://www.soc.surrey.ac.uk/JASSS/2/3/2.html)
[9] Edmonds, B. (2001) *Learning Appropriate Contexts.* In: Akman, V. et. al (eds.) Modelling and Using Context - CONTEXT 2001, Dundee, July, 2001. Lecture Notes in Artificial Intelligence, 2116:143-155.
[10] Edmonds, B. (2002) Exploring the Value of Prediction in an Artificial Stock Market. Workshop on Adaptive Behavior in Anticipatory Learning Systems 2002, Edinburgh, Scotland, August, 2002. (ABiALs 2002). Butz V. M., Sigaud, O. and Gérard, P. (eds.) Anticipatory Behavior in Adaptive Learning Systems. Springer, Lecture Notes in Artificial Intelligence, 2684:262-281.
[11] Edmonds, B. (2002) Learning and Exploiting Context in Agents. Proceedings of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Bologna, Italy, July 2002. ACM Press, 1231-1238.
[12] Edmonds, B. (2002) The Social Embedding of Intelligence - Towards producing a machine that could pass the Turing Test. CPM Report 02-95, MMU.
[13] Edmonds, B. (2006) The Emergence of Symbiotic Groups Resulting From Skill-Differentiation and Tags. Journal of Artificial Societies and Social Simulation, 9(1). (http://jasss.soc.surrey.ac.uk/9/1/10.html).
[14] Edmonds, B. (2008). Achieving Consensus Among Agents – an opinion-dynamics model. CPM Report CPM-08-186, MMU, Manchester, UK.
[15] Edmonds, B. and Hales, D. (2003) Replication, Replication and Replication - Some Hard Lessons from Model Alignment. Journal of Artificial Societies and Social Simulation 6(4) (http://jasss.soc.surrey.ac.uk/6/4/11.html)
[16] Edmonds, B. and Moss, S. (2001) The Importance of Representing Cognitive Processes in Multi-Agent Models, Invited paper at Artificial Neural Networks - ICANN'2001, Aug 21-25 2001, Vienna, Austria. Published in: Dorffner, G., Bischof, H. and Hornik, K. (eds.), Lecture Notes in Computer Science, 2130:759-766.
[17] Edmonds, B. and Norling, E. (2007) Integrating Learning and Inference in Multi-Agent Systems Using Cognitive Context. In Antunes, L. and Takadama, K. (Eds.) Multi-Agent-Based Simulation VII, 4442:142-155.
[18] Edmonds, B., Norling, E. and Hales, D. (2008, in press) Towards the Emergence of Social Structure. Computational and Mathematical Organization Theory.
[19] Granovetter, M. (1985) Economic-Action and Social-Structure – The Problem of Embeddedness. American Journal Of Sociology 91:481-510.
[20] Hales, D. (2001) Tag Based Co-operation in Artificial Societies. Ph.D. Thesis, Department of Computer Science, University of Essex, UK.
[21] Hales, D. (2004) From Selfish Nodes to Cooperative Networks – Emergent Link-based Incentives in Peer-to-Peer Networks. In proceedings of The Fourth IEEE International Conference on Peer-to-Peer Computing (p2p2004), 25-27 August 2004, Zurich, Switzerland. IEEE Computer Society Press.
[22] Hales, D. and Edmonds, B. (2003) Evolving Social Rationality for MAS using "Tags", In Rosenschein, J. S., et al. (eds.) Proceedings of the 2nd International Conference on Autonomous Agents and Multiagent Systems, Melbourne, July 2003 (AAMAS03), ACM Press, 497-503.
[23] Hales, D. and Edmonds, B. (2005) Applying a socially-inspired technique (tags) to improve cooperation in P2P Networks, IEEE Transactions in Systems, Man and Cybernetics, 35:385-395.
[24] Hardin, G. (1968) The Tragedy of the Commons, Science, 162:1243-1248.
[25] Holland, J. The Effect of Labels (Tags) on Social Interactions. SFI Working Paper 93-10-064. Santa Fe Institute, Santa Fe, NM. 1993.
[26] Kummer, H., Daston, L., Gigerenzer, G. and Silk, J. (1997). The social intelligence hypothesis. In Weingart et. al (eds.), Human by Nature: between biology and the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, 157-179.
[27] Palmer, R., Arthur, W.B., Holland, J.H., LeBaron, B., Taylor, P.(1994) Artificial economic life – a simple model of a stock market. Physica D 75:264–274
[28] Stolzmann, W. Anticipatory Classifier Systems. in Genetic Programming. 1998. University of Wisconsin, Madison, Wisconsin: Morgan Kaufmann.
[29] Turing, A.. M. (1936) On computable numbers, with an application to the Entscheidungsproblem. Proc. Lond. Math. Soc. 42:230-65; 43:544-6.