Gabriela Lindemann
Daniel Moldt
Mario Paolucci (Eds.)

Hot Topics

LNAI 2934

Regulated Agent-Based Social Systems

First International Workshop, RASTA 2002
Bologna, Italy, July 2002
Revised Selected and Invited Papers

Springer

Lecture Notes in Artificial Intelligence 2934

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Gabriela Lindemann Daniel Moldt
Mario Paolucci (Eds.)

Regulated Agent-Based Social Systems

First International Workshop, RASTA 2002
Bologna, Italy, July 16, 2002
Revised Selected and Invited Papers

Springer

eBook ISBN: 3-540-25867-1
Print ISBN: 3-540-20923-9

Created in the United States of America

Visit Springer's eBookstore at: http://ebooks.springerlink.com
and the Springer Global Website Online at: http://www.springeronline.com

Preface

This volume presents selected, extended and reviewed versions of the papers presented at the 1st International Workshop on Regulated Agent Systems: Theory and Applications (RASTA 2002), a workshop co-located with the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002), which was held in Bologna, Italy, in July, 2002. In addition, several new papers on the workshop theme appear here as the result of a further call for participation.

Agent-technology is the latest paradigm of software engineering methodology. The development of autonomous, mobile, and intelligent agents brings new challenges to the field. Agent technologies and multiagent systems are among the most vibrant and active research areas of computer science. At the same time commercial applications of agents are gaining attention. The construction of artificial (agent) societies leads to questions that already have been asked for human societies. Computer scientists have adopted terms like emerging behavior, self-organization, and evolutionary theory in an intuitive manner. Multiagent system researchers have started to develop agents with *social* abilities and complex *social* systems.

However, most of these systems lack the foundation of the *social sciences*. The intention of the RASTA workshop, and of this volume, is to bring together researchers from computer science as well as the social sciences who see their common interest in social theories for the construction and regulation of multiagent systems.

A total of 17 papers appear in this volume, out of 31 papers submitted. They include nine papers presented in the workshop (whose preproceeedings were published as *Communications Vol. 318 Mitteilung 318* of Hamburg University, Faculty of Informatics), as well as six new papers. In addition, an invited paper from Bruce Edmonds reflects some aspects of the lively discussions held during the workshop. The selection presented is divided into two major topics.

Topic A – *Social Theory for Agent Technology (Socionics)*

The wide range of social theories offers many different solutions to problems found in complex (computer) systems. Which theories, and how and when to apply them is a major challenge. In developing agents and multiagent systems computer scientists have used sociological terms like negotiation, interaction, contracts, agreement, organization, cohesion, social order, and collaboration. Meanwhile an interdisciplinary area called socionics, the bridge between sociology and computer science, is beginning to establish itself. The realization that the behavior of societies cannot fully be explained by macrotheories only, and the progress made in agent technology have opened the way to new models of societies in which both macrotheories and microtheories are incorporated. The development of the socionics research area and the increased interest in the dynamics of the behavior of agents in hybrid organizations requires the investigation of new modelling concepts like roles, groups, social intelligence, emotions, beliefs, desires, and intentions.

Topic B – *Norms and Institutions in MAS*

Multiagent systems are increasingly being considered a viable technological basis for implementing complex, open systems such as electronic marketplaces, virtual enterprises, political coalition support systems, etc. The design of open systems in such domains poses a number of difficult challenges, including the need to cope with unreliable communication and network infrastructures, the need to address incompatible assumptions and limited trust among independently developed agents, and the necessity to detect and respond to systemic failures.

Human organizations and societies have successfully coped with similar problems of coordination, cooperation, etc., in short, with the challenge of social order, mainly by developing norms and conventions, that is, specifications of behavior that all society members are expected to conform to, and that undergo efficient forms of decentralized control. In most societies, norms are backed by a variety of social institutions that enforce law and order (e.g., courts, police), monitor for and respond to emergencies (e.g. ambulance service), prevent and recover from unanticipated disasters (e.g., coast guard, firefighters), etc. In that way, civilized societies allow citizens to utilize relatively simple and efficient rules of behavior, offloading the prevention and recovery of many problem types to social institutions that can handle them efficiently and effectively by virtue of their economies of scale and widely accepted legitimacy. Successful civil societies have thus achieved a division of labor between individuals and institutions that decreases the “barriers to survival” for each citizen, while helping to increase the welfare of the society as a whole.

Several researchers have recognized that the design of open multiagent systems can benefit from abstractions analogous to those employed by our robust and relatively successful societies and organizations. There is a growing body of work that touches upon the concepts of norms and institutions in the context of multiagent systems.

July 2003

Daniel Moldt
Gabriela Lindemann
Mario Paolucci

Organization

The International Workshop on Regulated Agent-Based Social Systems: Theories and Applications (RASTA 2002) was organized by: *the Institute of Cognitive Sciences and Technologies* - CNR, Italy; *MIT Sloan School of Management,* USA; *AI Lab of the Department of Computer Sciences,* Humboldt University, Berlin; and *the Theoretical Foundations of Computer Science Group,* University of Hamburg.

Workshop Chairs

Daniel Moldt
Gabriela Lindemann
Mario Paolucci
Bin Yu

Organizing Committee

Rosaria Conte
Chris Dellarocas
Henry A. Kautz
Gabriela Lindemann
Daniel Moldt
Mario Paolucci
Munindar P. Singh
Bin Yu

Program Committee

Andreas Abecker
Karl Aberer
Mark S. Ackerman
Sven Brückner
Kathleen Carley
Jose Carmo
Enhong Chen
Helder Coelho
Rosaria Conte
Noshir Contractor
Raymond D'Amore
Kerstin Dautenhahn
Fiorella De Rosis
Chris Dellarocas
Frank Dignum
Peter Dittrich
Rino Falcone
David Hales
Andrea Hollingshead
Michael Huhns
Andrew Jones
Catholijn Jonker
Henry A. Kautz
Stefan Kirn

Victor Lesser
Ioan Alfred Letia
Henry Lieberman
Gabriela Lindemann
Jiming Liu
Steve Marsh
Mark Maybury
Ivica Mitrovic
Daniel Moldt
Bonnie Nardi
Hiroaki Ogata
Sascha Ossowski
Pietro Panzarasa
Mario Paolucci
Mirko Petric
Paolo Petta
Michael Prietula
Juan Antonio Rodriguez-Aguilar
Giovanni Sartor
Bernd Schmidt
Ingo Schulz-Schaeffer
Bart Selman
Carles Sierra
Munindar P. Singh
Sorin Solomon
Katia Sycara
Ingo Timm
Inga Tomic-Koludrovic
Adelinde Uhrmacher
Thomas Uthmann
Leon Van der Torre
Harko Verhagen
Pinar Yolum
Bin Yu

Referees *(not included in the Program Committee)*

Luis Antunes
Joscha Bach
Francois Bousquet
Jan Broersen
Marc Esteva
Eduardo Fermé
Guido Fioretti
David Hales
Xiaolong Jin
Michael Köhler
Maria Miceli
Dagmar Monett
Tim Norman
Alexander Osherenko
Giovanni Pezzulo
Heiko Rölke
Martijn Schut
Luca Tummolini

Table of Contents

Invited Paper

Topic A: Social Theory for Agent Technology

Topic B: Norms and Institutions in MAS

How Formal Logic Can Fail to Be Useful for Modelling or Designing MAS

Bruce Edmonds

Centre for Policy Modelling
Manchester Metropolitan University
http://cfpm.org/~bruce

"To a person with a hammer, every screw looks like a nail" (trad.)

Abstract. There is a certain style of paper which has become traditional in MAS – one where a formal logic is introduced to express some ideas, or where a logic is extended on the basis that it then covers certain particular cases, but where the logic is not actually *used* to make any substantial inferences and no application of the logic demonstrated. I argue that although these papers do follow a certain tradition, that they are not useful given the state of MAS and should, in future, be rejected as premature (just as if one had simulation but never run it). I counter the argument that theory is necessary by denying that the theory has to be so abstract. I counter the argument that logic helps communication on the simple grounds that for most people it doesn't. I argue that the type of logic that tends to be used in these papers is inappropriate. I finish with some suggestions as to useful ways forward.

1 Introduction

During RASTA 2002 there was some discussion about the utility of formal systems for building or understanding multi-agent systems (MAS). This paper is an attempt to put my arguments. I argue that (as with any tool) one has to use formal systems appropriately. Merely following a tradition of how to use and develop a particular kind of formal system is not sufficient to ensure one is doing something useful.

In this context I wish to make it clear that I have nothing against logic. I like formal logics because they can deal with qualitative information and they can be quite expressive. However, at the end of the day[1], they are just one of a range of types formal systems that could be used – the kind of the system that is chosen is important. The point is to distinguish when and how a particular formal system is useful – this applies to formal logics as a particular case.

In short, the question is not whether to abstract from our field of study using formal systems but how. In the past, premature 'armchair theorising' has not helped the eventual emergence of useful theory, but rather impeded it. Formal systems (such

[1] As David Hales would say.

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 1–15, 2004.

as logics) are not the *content* of theory but merely a *tool* for expressing and applying theory in a symbolic way – choosing the wrong kind of formal system will bias our attempts and make our task more difficult.

2 Two 20th Century Trends in Logic

Whitehead and Russell [13] showed that set theory, arithmetic and a good chunk of other mathematics could be formalised using first-order classical predicate logic. This dramatically demonstrated the expressive power of logic. Once set theory was properly logically formalised and the expressive power of set theory revealed it became clear that all mathematics could be embedded in set theory and hence be logically formalised. If any system could be shown to have an embedding in set theory, then it counted as mathematics. Thus set theory and classical first order predicate calculus was shown to *general systems,* in the sense that all known formal systems could be expressed in them (albeit with different degrees of difficulty).

In the second half of the 20th Century there was an explosion of different kinds of logic. This can be divided up into two approaches: those who were searching for the 'one true logic' (what I call the 'philosophical approach'); and those who saw logic as merely a useful tool for doing complex inference (what I call the 'pragmatic approach'). The former of these tinkered with the very structure of logic, restructuring the nature of deduction in the logic so as to attempt to match correct inference in natural language and by inventing new objects into the logic such as indices, operators, names etc. The nature of their discussions went very much by example – since they felt it was worth trying to construct the 'one true logic' it necessarily had to include all such cases. Logics in this vein included intuitionistic logic, free logic, relevance logic and modal logic. Due to the nature of their discussions their work tended to concentrate upon the axioms of the logic in relation to particular cases and treat the proof theory and formal semantics more as an after thought.

The pragmatic approach does not care so much about the philosophical interpretation as to what could be done with the logic. Thus, since classical first order predicate logic was generally expressive [7], they tended to work within this framework or construct simple extensions of it. For these people it was the pragmatic virtues that mattered: was it good for doing inference in; were its formal semantics checkable; was it easy to model with; and could it be used for computation (ala Prolog and its successors)? The particular logic chosen for the MAS modelling language, SDML[2] is a case in point – its purpose is not to capture any general theory of cognition but to provide a sound and efficient basic for the consistent firing of complex sets of interdependent rules [12].

Unfortunately the philosophical approach has tended to attract the more attention in AI. There may be many reasons for this: it may be that the association with philosophy gives it academic status; it may be that the participants truly believe that

[2] http://sdml.cfpm.org

there will be general logical systems that encode cognitive relations in ultimately simple ways; and it may be that it is relatively easy to write but difficult to criticise. Whatever the reason there has grown up a tradition in AI (and now MAS) which discusses different axiomatisations of logical systems based purely on plausibility and the ability to encode particular examples (i.e. its expressive power). It is this style of paper that I am arguing against on the grounds that, in the absence of any results, it does not merit publication.

3 Generality and Abstraction

One of the principle ways of achieving generality is to abstract away from the detail of particular cases leaving only what happens to be true of the wider domain one is considering (post hoc abstraction). Another way is to decide the structure before hand and to *choose* one's domain accordingly or else to simply ignore those aspects of those cases that seem to contradict that structure (a priori abstraction). A third way is to include a method for adapting to the particularities of each case so that the detail is preserved (adaptive generality). However it is achieved, the increased generality is obtained at a cost, a cost of lost information, relevance or computation respectively. The cost of losing information as a result of post hoc abstraction may be critical if it is the important details (w.r.t. one's goal) that are lost. The cost of restricted relevance as a result of a priori abstraction may be critical if this means that it excludes your intended object of study. The cost of increased computation may be critical if the computation is too onerous to be practical.

One well-known dynamic of philosophical discourse is that of the counter-example followed by an increase in generality: a thesis is proposed; then a case exhibited where the thesis fails; and, in response, the thesis is generalised (e.g. by adding caveats, or by being suitably elaborated). The repeated application of this process of a priori abstraction is a set of very general, but irrelevant principles. These principles may give one the illusion of relevance because the 'ghosts' of the original concepts are left as labels and symbols in the general principles and one has the impression that the relevance can be restored by the simple adding of particulars. However, if this attempted this is found to be unworkable in practice. Be clear – it is not generality or abstraction by themselves that causes this lack of relevance but the *way* the generality is achieved (i.e. a priori abstraction). Similarly – I am not arguing against generality or abstraction but that it should be done in a way that results in useful theory. Work which attempts to mimic the counter-example-generalisation process in formal logic will not result in relevant theory about MAS.

One way of clearly demonstrating that increased generality is not a sufficient reason for exhibiting a logic is that there are already many logics (and other formal systems) that are as general as possible. If a particular logic has the ability to capture a particular concept then the general one will also be able to do this. The point of inventing new formal systems is thus *entirely* pragmatic, for each system (even the general ones) will inevitably facilitate the construction of certain systems and frustrate others, just as different programming languages are good at certain tasks and bad at others. This presence of implicit bias is not a question of the theoretical ability of the system but practical ease for us humans. This is why we neither formalise everything in set theory nor program using Turing Machines. Choosing an inappropriate formal system will bias the development of a theory in unhelpful ways, choosing an appropriate system will facilitate it [4]. Merely establishing that a particular system can express certain properties does not demonstrate that the system will facilitate a good theory, for the general systems also do this and they would (almost certainly) make formal modelling impossibly cumbersome and inference infeasible.

Thus arguing for a particular kind of formal logic on the grounds that it is able to express certain ideas, concepts or cases is very weak, for there are already formal logics that do the same (if any can). Thus, although the development of formal logics is often driven by a wish to express certain ideas, they need to be *justified* on other, stronger grounds.

4 The Need for Theory

Clearly if we are to escape simply considering individual cases and if our understanding of MAS is to inform our construction of MAS (and *vice versa*) then we will need to generalise and abstract our knowledge, i.e. use 'theory'. The trouble is that 'theory' can come in a variety of levels of abstraction and a variety of forms. A natural language description is already a sort of theory because it is the result of many relevance and representational decisions – it provides a level of generalisation by facilitating the comparison of phenomena by substituting the comparison of descriptions. An MAS may be also be used as a method of producing a sort of dynamic description of a social system – this is when one attempts to program the individual agents as closely to actual accounts as possible and then check that all stages also correspond to those in the social systems at all levels of aggregation. Another MAS may be intended to represent a set of phenomena that occurs in a small set of individual cases – here the generality is restricted to a particular domain. At the other end of the scale are the 'high theories' of philosophy or sociology – these are ideas that are supposed to have a very great level of generality. In philosophy the theories tend to be precise but irrelevant. In contrast, in sociology the theories are relevant but often extremely difficult to pin down – they are more akin to a richly expressive language for talking and thinking about social phenomena.

I am unsure of exactly what Rosaria Conte means by 'theory' during her remarks during the closing panel of AAMAS 2002 (and elsewhere, see [2]). If she meant that *some* level of abstraction will be necessary for escaping from individual cases, then I agree with her – simply constructing particular MAS is not enough. However, if she is arguing that 'high theory' is necessary, then I disagree, for intermediate levels of abstraction also allows us to escape from single cases. For example physicists managed perfectly well to develop useful theories before the advent of their high theories, indeed they are still looking for a 'Theory of Everything' (TOE), even though it is clear that the situations in which such a TOE would diverge from the more mundane theories we already have will be extreme and unusual (from our point of view).

In the past theory that is mainly based on intuition which overtakes its evidential warrant has not had a good track record in resulting in useful theory. In fact, there is evidence that it has actively hindered the development of useful theory. A classic example of this is the thought of Aristotle on anatomy, which was wrong but played a part in delaying the spread of accurate information derived from dissection. Part of the reason for this is that theories play an important role in providing a language for thought, which (amongst other things) effects what evidence we look for [10] and biases further modelling effects (since other kinds of models will probably not fit well in that framework).

Thus papers proposing ‘high theories’ of MAS need substantial justification before being trusted and certainly more than a few cases and vague intuitions. Further, such high theory is unnecessary in order to escape particular cases and experiences – models that are specific to particular kinds of MAS and only somewhat abstract may be at a more appropriate level of abstraction and hence more reliable.

5 Different Stages of Science

If a particular language of thought is correct in the sense that its structure is itself well validated, then it might be well be profitable to explore. This is the situation that prevails in what Kuhn [10] called ‘normal science’ – a theory has been discovered and validated and then there is a stage of exploring the ramifications of this theory, applying the theory and using the theory as a means of guiding the search for new theories. This stage of science can be characterised as relatively cooperative and inward looking time – the participants tend to specialise into complementary skills and tasks and put these together within the established framework. There is a lot of ‘building’ on each other’s work and the field establishes norms so that new entrants to the field are required to strongly situate their contribution within the established framework, for example by citing those considered authorities. This can have the effect of excluding outside ideas so that the field becomes inward looking. In extreme cases this results in the ‘degenerate programmes’ described in [11].

During a period of normal science it may be sensible to simply accept the established principles, methods and assumptions and to concentrate on specialising and then developing complementary areas of knowledge using them. During such a time when those in the field are all using the same framework and outsiders are rare, one can take the common language of the participants for granted and simply use it as a vehicle for discussion.

Occasionally normal science is punctuated by periods of ‘revolutionary science’. This is when the established framework (if any) has become (or is revealed as) unsatisfactory and if a new and better framework is introduces it may become accepted. During such a period very little can be taken for granted, especially the assumptions and methodology of the old framework. Instead of cooperation and complementarily, sharp competition between different ideas and methods dominates.

Contributions are judged less by adherence to a particular framework and more by results. Typically in such periods one gets many contributions and academics from other fields being both offered and accepted.

During periods of revolutionary science one can not merely carry on with ‘business as usual’. Contributions to knowledge need to be more thoroughly justified in terms of results and (since there is likely to be a diverse audience) explained without assuming that all will understand the same language of expression. Since even the framework is in flux, what the relevant authorities for citing are unclear and it is not necessarily helpful to use established methods.

Neither the simulation of MAS nor their design has an established and well validated framework. There is no ‘high theory’ of MAS, and no proven methods. Whilst it is true that some people have *claimed* the status of authorities, whether posterity will agree will depend upon how useful their contributions turn out to be. A paper that might well be acceptable by those inside a field during a period of normal science can be found wanting in periods of revolutionary science, especially in the extent to which it justifies its method and proves its usefulness through its results. In [3] the relationship between formal systems and the dynamics of science is discussed in more detail.

A confusion about the stage that MAS is at may explain why some authors present their papers as they do – borrowing the style rather than the substantiality of papers in more successful sciences. If MAS did have a well validated general theoretical framework, then it might be more acceptable to present a exploration of part of that framework in a theoretical way, copying the methodology of accepted authorities in the field. Indeed, some of these papers *do* seem to imply that the use of simplistic deontic and epistemic logics *have* been established and proven, so what is left is to argue the details and make small extensions of these. Unfortunately this is far from the case – this style of formalism still has everything to prove.

6 What Sort of Logic Is Suited for Modelling MAS?

Since, in common with many other styles of formal system, logic has the possibility of modelling any system (via the truths concerning that system), it not so much a question of whether logic per se is or is not the correct kind of system, but more the particular type of logic that is used[3]. In particular it has tended to be the axiomatics of non-temporal, context-independent and propositional logics which are commonly discussed in this domain. This is in keeping with the philosophical logic tradition briefly discussed above. However, it seems patently clear that, if one is going to use formal logics in this domain, that it is the formal semantics of temporal, and contextual predicate logics that are far more appropriate. I consider these aspects in the following subsections.

[3] Although this still leaves question of the appropriateness of the implicit bias of the system.

6.1 Time

There are many ways of interpreting what logic *is* – as many ways as there are of interpreting the syntactic systems that constitute formal logic. Some see it as a way of defining a set of truths using inference or formal semantics, others see the inference as the most important which can be used for inference of conclusions (including the set of truths). Different people put the emphasis on different parts (which they may see as primary) and see the other parts as coming from these. However you see it logic relates a class of *truths* with a system of *inference*[4](embodied either in the proof theory as allowable steps or as the formal semantic validity of expressions expressing an implication).

As such it is hard to see how a logic can usefully model the connection between goals and actions without including an explicit representation of time. For example, the relation between the goal indicated by the utterance "I want to go for a walk tomorrow" and the present action of "cancelling a meeting scheduled for tomorrow" has an important temporal element to it. Yet almost all of the logics that have to do with goals and actions (including the deontic and BDI logics) do not have any *explicit* temporal element, instead they attempt to capture either the instantaneous or unchanging aspects in the relationship between such as: beliefs, desires, norms, goals, actions. In the first case they must miss the dynamic nature of the relationships, for example that one might change one's intentions as the result of weighing the effect of violating a social norm – indeed such an approach rules out any *interaction* between these entities at all. In the later (unchanging) case, one is limited to modelling only those aspects of the relationship that are always the case – thus if sometimes (but not always) a belief changes a desire and sometimes (but not always) a desire changes a belief then these relationships will not be universal over time. In this case it is an implicit assumption that the important relationships *are* abstractable without reference to temporal contingencies, which is extremely unlikely and without justification[5].

The other approach is to use implication as an implicit model of causation and thus encode the relevant sequencing in the axioms. The result of trying to fudge the issue in this manner is that the essential elements of the situation are represented by ludicrous propositions such as A = *the assertion that state of the world is such that I will be walking tomorrow* and B = *the assertion that I will take an action which I believe will prevent a future event which would imply* $\neg A$. This sort of move does nothing to convince me that this method of formalisation is capturing the essence of the case. Yet this is the case with many attempts which attempt to concertina concepts which are temporality situated into a non-temporal framework – representing *processes* as *single states* is bound to lead to huge practical difficulties if the framework was ever used for real problems.

[4] For a thorough discussion of the nature of logic see [7].

[5] I know of no attempts to justify such an assumption, rather the development of such logical formalisms seems to be on the basis that *any* caputuring of such mental entities is impressive and hence interesting, so it is felt that simple plausibility is sufficient to justify such explorations.

6.2 Lack of Formal Semantics

Another strange fact about the style of formal logic that have been discussed in RASTA and more generally in MAS is the lack of formal semantics. If one is primarily concerned with the *meaning* of modal operators and determining which ones are *valid* then the formal semantics are much more relevant than the axioms and proof theory. A logic that had as its universe of models (models in the logical sense) a set of MAS outcomes (i.e. the set of possible MAS states over time) and showed that certain expressions were logically validated w.r.t. these semantics, would be a useful development. On the other hand if one is more interested in inference (being able to infer conclusions from premises) then the proof theory is more important (in this latter case, one would expect minimal discussion of the meaning of operators and a focussing on the useful and interesting inferences that can be obtained using the proof theory).

6.3 Context Dependency

The typical presentation of logic in MAS assumes and depends upon the fact that all the reasoning is done within a single context. Sometimes this is explicit, but more often it is left implicit and only indicated by the test problems (if any). This is very strange because reasoning about norms, goals, intentions, learning is only feasible if one can relate these to the contexts, for example intentions may involve action in several different contexts or involve explicitly effecting what the context is.

Whilst taking the context-dependency of many of these concepts seriously does mean accepting that it will be difficult to generalise, the pay-off id that context-dependent reasoning (and learning) is far more pratical and feasible that the general variety.

6.4 Numbers

A final area I will deal with is the ability of logic for understanding or designing MAS that does not allow for an adequate arithmetic. MAS in which numbers play no significant part are hard to find, but despite this most of the logics proposed rule out any sort of predicate logic in which such numbers could be defined. The reason for this is, presumably, because the introduction of arithmetic means that there can be no complete formalisation of truth, that is to say that there will be no method of proof that will be able to prove all the truths. This is due to Gödel's incompleteness theorem. However, the goal of completeness is simply inappropriate for almost all MAS – we are never going to be able to prove all an MAS's properties. Thus eliminating numbers to retain completeness is not sensible – it is a case of changing the problem to suit the tool.

Of course, a temporal contextual predicate logic with semantics that can capture multi-agent belief will not be such a clean simple system as those frequently discussed, but that is appropriate because most MAS are not clean simple systems! In this case simplicity is certainly not indicative of usefulness, let alone truth [5]. Some will argue is that they are deliberately abstracting away from the detail of time, context, and numbers in order to obtain a general theory, but the burden of proof is then surely upon them to show that they have done this successfully. Justifying such extreme abstraction on the basis of a few intuitions does not wash – the wish for the 'magic' shortcut is strong but can not be relied upon.

On the other hand, if proponents of such formalisms tried to use their constructions on real problems or to model real systems, the inadequacies and over-simplicity would be quickly revealed. If (as I suspect) there were no adequate work around that preserved the logic then this would be revealed and if there were it would be demonstrated how and in what way this formalism would work.

7 The Audience's Viewpoint

When presenting results there is an understandable wish on the part of the authors to concentrate on what they have *done*. However, for the audience it is more important to first of all judge whether the work is worth learning about or even applying. This is because they are bombarded with ideas people have had and systems they have designed – they are not short of ideas, but they do need help in deciding which ideas or systems to invest their time and effort in. Everybody feels convinced that their ideas or systems will work, otherwise they would not be presenting them. Similarly, everybody has some sort of thought train that lead them along the path they took, so everybody has *some* good reasons for doing what they did. Thus the presence of good reasons for doing something does not help an audience distinguish between different ideas or systems, more is needed.

One claim for formal logic made during the discussion at RASTA 2002 was that it aids communication because it allows one to be precise about ideas. That they allow one to be precise is true, formal systems (even if totally misguided) at least make for an unambiguous common referent. This is particularly attractive for disciplines which are bedevilled by different approaches, vagueness and misunderstandings with respect to their key terms. Precision is definitely a virtue, but it is not sufficient to ensure good communication. Whether formal logic does or does not aid communication is an empirical matter. Frankly, I doubt whether this was true for the audience we had at RASTA, for these logic papers are only accessible to the small minority who had sufficient familiarity with formal logic to be able to fluently 'read' it.

Even if there we assume that formal logic *did* aid communication between those who had suitable training, this still is insufficient to justify such a presentation. Being crystal clear in one's communication is no good if what is being communicated is not worth the effort. What was being communicated in some of these papers was simply unproved ideas and intuitions – directly comparable to specifications for systems that have not been implemented or otherwise tested.

Further the fact that the ideas and intuitions were expressed using formal logical expressions served to prevent the majority of the audience from evaluating them, leaving this evaluation to an 'in crowd' who are, on the whole, already sympathetic to the approach. It is almost certain that if I had not been there (being a person who is both critical and sufficiently knowledgeable of formal logic) there would have been no discussion about the worth of the formal logical approaches presented. Now I am sure that it was not the intention of the formalists to use their formalisms as a way of preventing criticism or ensuring acceptance, but this would have been the effect.

Thus a paper which does not provide any evidence for the usefulness of a formalism (apart from the reasons that lead the authors to invent or extend it) simply fails to satisfy the justified norms of scientific communication because it ignores the needs of the audience to evaluate the suggestions. Further, a formal system that has been used for solving a real problem or modelling a realistically scaled MAS will be greatly improved and be more likely to introduce genuinely new ideas. Intuitions are highly biased by the current *Zeitgeist* which is why rubbing them against a real problem is more likely to provide new input than simply more discussion between other academics immersed in the same *Zeitgeist.*

8 A Common Argument for Formalism

However, a logician (or mathematician or whatever) may object in the following manner: "the history of the development of formal systems has included many systems that would have failed on your criteria and yet turned out to be immensely useful later - are you not in danger of arguing against similar advances with such warnings?" My answer is fourfold.

- Earlier, we did not have the huge number of formal systems we have today, and in particular we did not have the general systems mentioned above. Today we are overwhelmed by choice in respect to formal systems – unless substantial advances are made in their organization all new systems will need to be substantially justified if their clutter is not to overwhelm us.
- There are proper domains for formal systems that are purely conceptual: philosophy or pure mathematics. Presenting a formal system elsewhere implies that it is relevant to the people in the domain in which it is being presented. If it really is relevant to them this needs to be demonstrated.
- Even in pure mathematics presentations or publications are required to justify themselves appropriate criteria - novelty, expressiveness and soundness are not enough (although the other criteria perform a weaker role than when they are applied elsewhere). For example, in the examination of a doctoral thesis in pure mathematics once the soundness of the work is deemed acceptable it is the importance, generality and relevance of the results that are discussed.
- The cost structure of the modelling enterprise has changed with the advent of cheap computational power. It used to be the case that it was expensive in both time and other resources to use and apply a formal theory, so that it was important to restrict which formalisms were available. Given that the extensive validation of the success of formal systems was impossible they had to be selected almost entirely on a priori grounds. Only in the fullness of time was it possible to judge their more general ease of use or utility of their conclusions. Now this situation has

changed, so that the direct validational assessment of a formal system can be achieved with relative ease for relevant cases.

One can choose to judge a formal system by the criteria of pure mathematics (or logic) that is show the system has generality and inferential power by exhibiting theorems and proofs. One can choose to judge it as applied mathematics, whose criteria include problem solving ability and relevance by demonstrating its *use* in modelling systems. What is not acceptable is to fail to demonstrate that it succeeds by any kind of criteria. Some of the formalist papers in MAS fail in precisely this way, they excuse themselves of solving particular problems but also fail to exhibit and substantial theorems and proofs.

9 Some Suggestions for the Way Forward

It should be clear that I am not against the use of formal logics as a tool for understanding MAS *per se,* but against using them in unhelpful ways, namely as a language for philosophical discussion. Intuitions that are relatively unconstrained and unvalidated have a poor track-record when it comes to real applications and problems, and formalising these in relatively simple (and, I argued, inappropriate) logics does nothing to solve this basic problem. Simply following the form of a philosophical tradition is insufficient to justify the presentation of work – an audience rightly expects some conclusions in the form of results by which they can evaluate the ideas. Yet we do need somewhat abstract and precise models to improve our understanding, and logics are an expressive and flexible kind of formal system. So might be the way forward?

Before suggesting some steps we might take, I will describe our domain as I guess it is. I think that the study of social systems in general, and MAS in particular, will be more akin to biology than to physics and the production of MAS closer to stock breeding and ecological management than to traditional engineering[6]. I think that there will be hundreds of essentially different 'species' of MAS, all of which will need to be individually described studied rather than their being adequately covered by any easily accessible universal principles[7]. I think that there will not be any easy 'short cut' to useful high theory, and certainly not via vague intuitions expressed in formal logic. Thus I would the following (incomplete list) based upon analogies with other sciences:

- The development of new ways of collecting data and observing MAS;
- A considerable period of descriptive modelling (i.e. less abstract modelling) so that we have a way to compare different MAS;

[6] Or, at least, to a traditional account of what traditional engineers do. Engineers, in practice, don't actually act as these accounts would suggest. A classic example of this is the neatly ordered elicit; analyse; design; implement; test cycle that software engineers are supposed to follow.

[7] After all a 2D cellural automata with extremely simple rules for each node can implement a full Turing machine [9] and hence, in principle, *any* computation.

- A building up of complete chains made up of models at different levels of abstraction so that each are each clearly related (or relatable) to less abstract models;
- The insistence that any abstract or formal theory is treated with scepticism until it proves its worth – the more abstract it is the more is has to prove;
- That, nonetheless, we continue to try to build models whose level of abstraction is justified by (and judged by) the evidence;
- The rejection of papers that merely specify things based on single cases, intuitions and expressiveness because they are premature;
- That, nonetheless, the greatest variety of formal systems should be encouraged as possible members of a 'tool box' for MAS practioners and studiers (but only accepted after they have shown to be helpful in at least one real case).;
- That papers suggesting formal systems for helping design MAS should demonstrate that it is feasible to design an MAS that works using them;
- That papers suggesting formal systems for understanding MAS should show that they do, in fact, capture the phenomena they claim providing either a successful prediction or a credible explanation of that phenomena;
- That the field resists the temptation to retreat into formalism and philosophy when it substantial progress is difficult.

This is a more *pragmatic* and less ambitious approach than many academics have hoped for or will accept. They will continue to dream of inventing the 'magic bullet' that allows us to shortcut the large amount of messy empirical work that will be necessary and take us straight to powerful high theory (as, indeed, do I in moments of weakness). However I think this has more chance of producing useful knowledge and, *eventually,* useful theory. We will always continue to need some sort of abstractions to help us search, but until we have some well validated examples we need to stay as flexible as possible and stay suspicious of easy or prevalent intuitions.

10 An Exercise

Look through some of the papers in this (and similar) volumes. Does the 'conclusion' state what was done and why it was done but not state any results or conclusions (other than that the authors think it is the right way to do it)? Is there any way of *evaluating* what was done using the information in the paper? Is there any way of knowing when the techniques or ideas described in the paper would be useful to apply and when not? Have you been informed of anything except the present state of thought of the authors? If there were no results, did the system (either formal or software) help demonstrate or communicate the authors ideas effectively? Where those ideas so good to warrant presentation with no results?

One way of stripping bare the impressive effect that a formal logic imparts is to imagine the same sort of paper but using a simulation instead of a logic. If the paper was one where a simulation was described along with the reasons why it was so designed, but the simulation was not actually run and no results were shown, would it make a satisfactory paper? I think not.

11 Positive Examples of the Use of Logic in MAS

From the above it should be clear that I do *not* think that *all* work in MAS that employs formal logic is useless. It is inevitable (and often helpful) to abstract when trying to solve problems and develop techniques. If someone can achieve demonstrable results using such an abstract system (including formal logic), then this is undeniably useful and worthy of publication and attention.

For example Frank Dignum and his team often use formal systems as an approach to solving real world problems and developing systems to work in the real world (e.g. [1]). This works both at the conceptual and the computational level. 'Rubbing' formal systems 'against' real problems and domains can lead to interested and relevant lessons being learned.

12 Conclusion

In many ways Frank Dignum is an ideal person to answer my criticism of empty formalist papers. He does use logic in much of his work, but he applies these ideas in real implementations which are attempting to solve real problems. In my view it is exactly this "rubbing together" of abstract ideas and real domains which gives interest and relevance. He has obviously been inspired and aided by his study of formal logics. In their reply to me Dignum and Soneberg give several examples of this sort of inspiration and conclude that developing logical systems has, at times, been helpful in the design of MAS.

I certainly agree that abstraction and formal models (including logic-based ones) can be very helpful in solving practical implementation and modelling problems in MAS. Indeed I would argue that abstraction and formalisation are often essential if substantial progress is to be made. Further, my reading of Dignum and Soneberg's reply indicates that they also deplore the presentation of empty papers which do not present any results, or even implementations. So wherein lies the disagreement?

I think the difference lies in our views of the scientic process in which MAS is embedded. It appears that Dignum and Soneberg see empty formalist papers as an inevitable phenomenum – a sort of irritating, but ultimately irrelevant, "background noise". I, on the other hand, see this as a more active, detrimental and preventable phenomenum, which is why I bother to argue against them. Now the progress of MAS and the influence of particular papers and approaches is a very complex and varied affair – one is never going to be able to finally demonstrate which view is correct. However, I do think that an examination of such processes can be helpful in that some guides for future action can be made.

Let us consider the case of BDI logics and their ilk and their influence on the MAS community. It is undoubtabley the case that BDI logics have enjoyed a wave of popularity in recent years (fortunately now on the wane). It is also clear that their popularity during this time was not based upon any substantial evidence that their use provided any significant "leverage" for solving any real world problems. In particular it was not the case that the properties of the BDI logics were shown to be pivotal to the advertised BDI languages (e.g. dMars) or applications. Rather their use seemed to be as a sort of loose analogy for guiding programming. Their popularity seemed to be more based upon the vision of agent-based software engineering that accompanied them[8]. What resulted was that many papers were written to look like they were about or used BDI agents, when, in reality, they were not. Was this a case of simple and harmless "background noise" or did it, in fact, waste a lot of time of many researchers across the world? I leave the reader to decide.

In his examples Dignum points to a more productive way forward. Learn about and know a whole range of formal systems, so that when you are presented with a difficult problem you have a substantial palette of formal systems with which to solve it with. Providing this pallette is, indeed useful – it is what pure mathematics does. However this does not excuse publically presented papers of meeting some hard criteria – it is just that different criteria apply. A pure formalist paper needs to demonstrate its generality, potential relevance and inferential power [3]. If the empty formalist papers met these criteria I would not be complaining.

Dignum and Soneberg propose the slogan *"No experimentation without explanation",* meaning *"No published experimentation without an abstracted explanation"* (since an explanation which was simply a long trace of the particular computation does not help). I agree with this, however I add "*No published abstraction without results*". Together they form the criteria that you need both experimentation/results and explanation/abstraction for something to be worthy of presentation in a public forum (what consenting researchers do behind closed doors is, of course, their own affair).

I will end by describing a happy outcome of this interchange, and thus attempt to assuage Dignum's fears: that researchers will pay a little more attention to the needs of their audience when presenting formal systems and that reviewers will not be scared by heavy formalisms and be a little more strict at rejecting papers that show neither results nor demonstrate their inferential power. I think the outcome of this would not be to split the field or to stop the interchange of ideas, but cause the formalist papers that are presented to have more impact and become more productive.

References

1. Bons, R. Dignum, F., Lee, R. and Tan, Y.-H. A Formal Analysis of Auditing Principles for Electronic Trade Procedures. *International Journal of Electronic Commerce,* **5**(1):57–82 (2000).

[8] A readable and authoritative account of this vision can be found in [14] which I review in [4], discussing many of these issues.

2. Conte, R., Edmonds, B., Moss, S. and Sawyer, R. K. (2001). Sociology and Social Theory in Agent Based Social Simulation: A Symposium. *Computational and Mathematical Organization Theory,* **7**:183–205.
3. Edmonds, B. The Purpose and Place of Formal Systems in the Development of Science, CPM Report 00-75, MMU, UK (2000). `http://cfpm.org/cpmrep75.html`
4. Edmonds, B.. A review of Reasoning about Rational Agents. Journal of Artificial Societies and Social Simulation,5(1) (2000), <http://jasss.soc.surrey.ac.uk/5/1/reviews/edmonds.html>
5. Edmonds, B. (2002) *Simplicity is Not Truth-Indicative.* CPM Report 02-00, MMU, 2002. `http://cfpm.org/cpmrep99.html`
6. Gabbay, D. M. *Classical vs. non-classical logics : the universality of classic logic.* Saarbrücken : Max-Planck-Instituit für Informatik (1993).
7. Gabbay, D. M. (ed.) *What is a logical system?* Oxford: Clarendon Press (1994).
8. Gärdenfors, P. (1988) *Knowledge in flux : modeling the dynamics of epistemic states.* Cambridge, MA; MIT Press.
9. Hopcroft, J. E. and Ullman, J. D. *Introduction to automata theory, languages, and computation.* Boston: Addison-Wesley (2001).
10. Khun, T. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press (1962).
11. Lakatos, I. *The methodology of scientific research programmes.* Cambridge: Cambridge University Press (1983).
12. Moss, S., Gaylard, H., Wallis, S. and Edmonds, B.. SDML: A Multi-Agent Language for Organizational Modelling. *Computational and Mathematical Organization Theory,* **4**:43–69 (1998). `http://cfpm.org/cpmrep16.html`
13. Whitehead, A. N. and Russell, B. *Principia mathematica.* Cambridge: Cambridge University Press (1962, originally published 1913).
14. Wooldridge, M. Reasoning about Rational Agents, Cambridge, MA: MIT Press (2000).

Communicational Patterns as Basis of Organizational Structures

Steffen Albrecht and Maren Lübcke

Technical University of Hamburg-Harburg, Department of Technology Assessment,
Schwarzenbergstr. 95, 21071 Hamburg, Germany
{steffen.albrecht, maren.luebcke}@tu-harburg.de

Abstract. Researchers in Distributed Artificial Intelligence have employed the notion of "organization" to guide the design of distributed software systems. There is a growing consciousness that MAS designers have to be aware of the social factors underlying the formal organizational design. The study presented in this paper attempts to contribute to this development threefold: on a conceptual level, we offer a notion of organizational structures grounded in the theory of social systems according to Niklas Luhmann. On a methodological level, we employ methods of social network analysis as a tool for the detection and operationalization of such structures. Empirically, we demonstrate what results can be obtained by this approach to the observation of communicational patterns. With this study, we exemplify the fruitfulness and the scope of the novel perspective on organization for the design of MAS.

1 Introduction

Designers of multi-agent systems (MAS) share the paradigmatic view that using agents as abstract computational elements for the design of computer systems has a number of advantages which other methods of software engineering seem to lack. Agents are conceptualized as autonomous, intelligent, pro-active and socially interacting entities [34]. However, the decision to build systems based on such autonomous elements in turn confronts the designers of MAS with the problem of coordination [3]: How can agents interact most effectively to solve the tasks they were designed for? This question becomes even more problematic once we consider "large-scale open systems" [13], where the type of problems to be solved and the type of agents interacting is not known beforehand, but depends on the dynamic evolution of the system.

Since the early days of the discipline, researchers in Distributed Artificial Intelligence (DAI) have employed the notion of "organization" to guide the design of distributed software systems (cf. [7], [10] for an overview). Different to the use of other metaphors, the term "organization" was not only a source of inspiration, but entered the field of DAI together with a whole body of literature on organization

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 16–30, 2004.

theory and organization studies that had been developed in social sciences.[1] Organizational theory and studies have not led to a definite answer to the question of how to organize agents' interactions.[2] But it became clear that this question cannot be answered without the insights of the social sciences [12].

The study presented in this paper attempts to contribute to the research in this direction. We follow the basic idea of "socionics" [23] in that we exploit sociological insights from social and organization theory to support MAS developers. More specifically, we aim at enriching the perspective of MAS designers with a new way of thinking about and modeling organizational structures.

Up to now, researchers in DAI mainly borrowed sociological insights from symbolic interactionism (cf. [30]). However, one of the most prominent works in social theory – the "Theory of Social Systems" as it was developed by Niklas Luhmann [20], [21] – was very little referred to. Though genuinely interested in abstract social theory, Luhmann was also much concerned with organizations as one of the most outstanding phenomena in society [22]. His social theoretic perspective includes a radically new view on sociality itself, which is of great importance for conceptualizing the social aspects of MAS: While traditional social theories focus on the actions of agents,[3] systems theory emphasizes that communication has to be seen as the basis of social systems. Social systems consist of communication; all else (including the individual agents) is considered as belonging to the environment of the system. As a result of their operations, social systems establish boundaries towards their environment. Consequently, they have no direct influence on their environment.

This new paradigm offers a precise conception of sociality that is independent of assumptions about individual agents. With systems theory, we can reformulate the problem of coordination in MAS: How can social systems develop stable structures (which are necessary, e.g., to solve problems) in spite of a rapidly changing environment, in which agents come and go and over which the system has no direct control. Additionally, systems theory offers theoretical insights into organizational phenomena that allow to address this question in a fruitful way: Organizations are specific social systems. As a result of the evolution of society, they achieve a high degree of stability in complex and dynamic environments [20]. Thus, it seems promising for MAS designers to observe the mechanisms which foster stability in organizations.

In the remainder of this paper, we demonstrate how MAS designers can profit from the perspective on communication offered by systems theory. We discuss the mechanisms that allow organizations to operate continuously, and we present concepts to make these mechanisms observable. We propose to use methods of Social

[1] This "theory import" can be traced back to Fox' early work [11], in which he draws on the organization theory of Nobel-prize winner Herbert Simon.

[2] As Carley and Gasser note: "There is no single organizational design that yields the optimal performance under all conditions. Which organizational design is optimal depends on a variety of factors (...)" [7].

[3] Social theories mostly use the term "actors" instead of "agents". We use "agents" for both, humans and computational elements, to make it easier to reason about computer systems in the terms of social theory.

Network Analysis (SNA) to empirically observe and analyze communication in an organizational context. And we apply our approach in a case study of real-world organizational communication to show what kind of results can be expected from it and what value these have for the design of MAS.

Thus, our contribution to the problem of organizing interacting agents is threefold:

- On a *conceptual* level, we offer a notion of organization that is grounded in the theory of social systems according to Luhmann. With its roots in a social theory with universalistic scope, it provides a very broad and inspiring, yet precisely formulated theoretical framework to address questions of organizational design.
- On a *methodological* level, we propose to use SNA as a tool for the detection and operationalization of social structures. SNA seems promising since it has a well-formulated mathematical foundation in graph theory. Furthermore, applications to communication and organizational phenomena have proved to yield interesting results (cf. [2], [14], [26], [28], [32]). Applications in the field of DAI are diverse, e.g., besides our approach, Yu and Singh introduce SNA to find experts in a referral network [35], and Sabater and Sierra derive reputation measures from SNA [29].
- *Empirically,* we observe and analyze processes of communication in an organizational setting to derive knowledge about the structures that evolve and to evaluate the methodology. With our case study, we do not claim to achieve representative insights, but it serves to exemplify the fruitfulness and the scope of the perspective on organization we present here.

2 Organizations and Organizational Structures

Organizational design in MAS is mostly seen as a top-down formal specification of what tasks should be solved by whom in what way [15]. In this view, the concept of rules is important, since for an engineer, this seems to be the "lever" with which to implement the desired behavior into an agent. This position is taken for example in [36]. To enrich the perspective on computational organizations, we argue for a bottom-up approach that focuses on emergent properties rather than formal rules that have been pre-designed by the designer of a MAS.[4] In this, we follow researchers like Carley [6], who advices to theorize "from the ground up", and Ackerman and Halverson [1], who emphasize that "organizations are hardly a single, unified entity, as the metaphor implies".

There are a number of reasons to follow such a bottom-up approach, especially with respect to large and dynamic MAS: A designer might not know or might not foresee all tasks the agents in a MAS should fulfill. Thus, agents should be able to

[4] Such an understanding is also reflected by new concepts in business like "Process Reengineering" or "Total Quality Management" which combine top-down approaches with bottom-up strategies: The goal of the enterprise is defined by the management (top-down), but the way these targets have to be fulfilled are decided bottom-up with respect to the actual state of the environment to improve adaptability and proof correction.

alter their behavior according to future needs. The system's environment may change, so that the agents need to adapt their behavior and social structure. Such a capability can only be supported if the designer is aware of the emergent properties of social systems. Finally, if we design open MAS like for example personal assistants interacting on the Internet, we cannot know precisely what types of agents will together form the organization, and rather rely on emergent regularities than on pre-defined behaviors.

Such an approach is supported by Luhmann's system theory. Since it would be beyond the scope of this article to refer the whole underlying social theory, we focus on the main aspects that make this theory different from others, and we introduce the relevant concepts with respect to organizational analysis.

- *Communication* is the basic element of social systems. Every social system is organized autopoetically, which means that only the system itself is able to specify and change its structures (this is in contrast to, e.g., [7], where organizational structures can be formed by the system designer). The consequence is that agents belong to the environment of the system and are treated like "black boxes", with no opportunity to know precisely their internal state.
- As mentioned above, organizations are regarded as social systems that achieve a high degree of stability in dynamically changing environments. To achieve this robustness, they have developed *organizational structures.* Structures are mechanisms that reduce the complexity of the range of possible communication, i.e. they focus the operations of the organization. From the perspective of communication as the basis of social systems, structures are stabilized patterns of communication that evolve in the course of time. Luhmann [22] identifies three such mechanisms that reduce the level of variance in organizational operations: persons and their roles, communication channels, and programs.[5]
- *Persons* are communicative constructs that relate the social system with an individual agent. This construct helps to ascribe a number of utterances to one author. Thus, persons operate as addresses in the flow of communication, and they allow observers to build expectations about their behavior. The concept of persons allows to abstract from singular events towards long-term behavior. *Roles* go one step further; they allow to expect a behavior independent of a specific person. A role is characterized by a typical behavior. Individual agents can play a role, or external observers can observe it as characteristic for certain agents. Roles stabilize the operation of organizations by offering abstract models for individual behavior.
- *Communication channels* reduce complexity since they work as filters in the dynamic flow of communication. An organization considers only those messages as relevant that pass the official channels. The design of these channels can be changed by the organization according to the needs, e.g., centralized, hierarchical communication channels can be replaced by a more network-like structure.

[5] According to Luhmann, organizations typically communicate with the help of decisions. To keep our discussion on a more general level, we speak of communication in general and not specifically of decisions.

- *Programs* finally are for example oriented towards the goals of the organization. Organizations establish programs to filter out any operations that are not relevant for achieving the overall goals of the organization. Conditional programs are another example; they specify the reactions of the organization to events in the environment. In the following, we will not consider this type of organizational structure, since it is most closely related to how DAI researchers already conceptualize organizations.

In a typically MAS design, the implemented organizational structure defines the different roles of agents within the system, and these in turn determine the way agents can communicate with each other. In our approach, what is communicated and how it is communicated defines the whole organization. Communication constitutes the organizational structure and makes it observable. Our empirical analysis below illustrates this correlation. Persons, roles and communication channels can thus be reconstructed from the dynamic processes of communication, and this information can be fed back into the system to enrich its social intelligence.

This new view we offer is valuable especially with respect to very complex or highly dynamic MAS, in which unintended consequences occur and in which the fluctuation among participants is potentially very high.[6] In open systems, like on the Internet, we do not know all types of agents that might wish to interact. Thus, a perspective that conceives individual agents as “black boxes” is advantageous, since it draws attention to the social system layer that allows such agents to influence the communication in the system. With communication as the basic element of social systems, the theory is based on principally observable behavior and does not need to make assumptions about the internal states of the agents.

3 An Empirical Case Study

According to Luhmann, complex social systems, like, e.g., organizations, are not able to connect all their elements with each other, but have to design specific patterns of relations between them. This process of selection results in a specific structure, consisting of patterns of communication. To observe such communicational patterns, we propose to use methods of social network analysis. Research in SNA focuses on relational data (based on relationships between elements) in contrast to classical methods dealing with attributional data (for a comprehensive overview, see [31]). SNA makes only few theoretical assumptions, but it offers a wide range of mathematical methods to describe and analyze social relationships.

To demonstrate the potential of the systems theoretical approach in the context of organizations, we apply the basic concepts outlined above to a real world example of intra-organizational communication. The data come from a moderated debate among

[6] In this respect, the theory of social systems has much in common with the organization theory of March [24], [25], who is concerned with exactly such problems.

the members of a university. The participants, students and faculty members, engaged in an online discussion about the quality of teaching.[7]

The main objective of this discussion was to identify the problems of the actual evaluation of lectures and to work out improvements for the evaluation program. About 1010 students of three faculties were asked to discuss this topic with their 52 professors for six weeks in November and December 2001. More than 20% of this target group registered (n=228 contributing participants), and another half of them participated actively in the discussion, 70% with two or more contributions. 1211 messages were posted in the debate, an average of about 30 contributions per day. The debate was moderated by a team of researchers who structured the dialogue and continuously summarized its results.

We consider the process of communication in this debate (i.e. the exchange of utterances in a "new contribution" / "reply to" – structure) as the observable part of the underlying social system. The contributions to the debate and the structure of the discourse allow to observe a facet of the organization "in actu", and the measures of SNA give us the means to analyze its structure. The data itself consists of the anonymous participants' contributions to the debate. We employ the notion of 'person' as communicative construct that operates as an address for building expectations. The (potentially virtual) identities of the users (i.e. user names) form such addresses, to which we can ascribe a number of contributions. Thus, persons form the nodes of the network. As usual in online communication, the contributions were ordered in a threaded structure, i.e. relations between contributions are of the type "reply to ...". A contribution connects its author with the person who wrote the contribution to which it refers. With persons as nodes and referrals as ties, a graph of the communicative network can be obtained easily.

Our aim now is to show how methods of SNA can be used to observe a social system. Building on persons and their communicative relations, we analyze the overall structure of the communicative exchange as well as typical patterns that emerge in the course of time. Structures in the sociological sense evolve from the bottom up, once specific patterns are repeated and reinforced. Such redundancies can stabilize the otherwise chaotic behavior of social systems, and they can serve for an observer as the basis for predictions of the system's future behavior. If a system is redundant, an external observer needs only little information to estimate the behavior of the whole system. Such observations in turn help to reinforce the emerging structures, and increase the stability of the social system.

We are especially interested in two mechanisms that enhance the redundancy in organizations: roles as abstractions from individual behavior, and communication channels as abstractions from particular opportunities for communication. If an observer knows what role a person has in the organization, and if he knows what communication channels are used by persons employing such a role, he can use this

[7] The online forum system used – "DEMOS – Delphi Mediation Online System" – was conceptualized at our department in the context of the project "DEMOS" (partially funded by the European Commission under the IST program. For further information about DEMOS, see [18], [19], or the project's website: www.demos-project.org.

knowledge to build expectations about this person's future reactions [20]. Our empirical analysis seeks to exemplify the powerful information that can be derived from an analysis of communicative patterns, and how it can be used by autonomous agents to monitor and cope with dynamic and complex social environments. With this focus, we follow ideas by Rauch [27], who argues that large-scale discourses can serve as a kind of "burning glass" for the surrounding social system. We also follow Lorentzen, Nickles, Weiß and Brauer [17], [4], who employ a so-called "social system mirror" concept to enhance MAS with reflexive social knowledge.

4 Observing Communicational Patterns

Our analysis of the communicative network in the debate focuses on three aspects of the graph structure: a global view on the topology of the network, an analysis of the different roles agents typically played, and an analysis of the communication channels that were used in different parts of the network.

The global topology of the network of communications makes the top-level structure of the social system observable. It can be analyzed with the help of a decomposition of the network. For each separate level of tie strength, we analyzed which components (i.e. connected groups of agents that are not related by any ties to other groups or individual actors in the network) can be distinguished in the network. To visualize this analysis, we mapped the component structure onto the different levels of tie strength, as can be seen in figure 1.

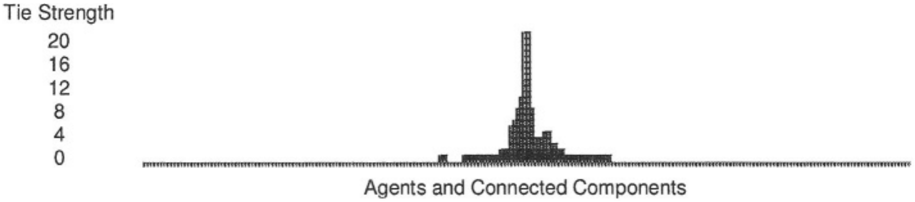


Fig. 1. Mapping of the overall component structure of the communication network. Each column equals one agent (n=228). On each level of tie strength, horizontally neighbored agents belong to the same component

The structure obtained from the mapping resembles a hill with mainly one central peak. This central peak indicates a strongly interconnected core group of agents (maximum tie strength greater than 20). Towards the bottom of the hill, the basis gets larger, including a greater number of agents in the central component. This strong main component indicates that on a basic level, most agents are connected with each other via communication. The connections are mediated by the strongly connected core group of agents that function as integrators and connectors. Thus, we can observe from the communications a homogeneous, integrative structure.

This observation is confirmed by a second property of the mapping: there exist only very few divergences between agent components in the structure. The diversity of the peaks represents the diversification of the organization. For example, in a highly differentiated organization, we would expect to observe a structure with deep gaps between the different divisions. In our case, however, we can observe only two small peaks diverging from the central one, indicating a very high interconnectedness between all individual agents.

Thus, our component mapping gives an overall visual representation of the underlying organizational structure, operationalized in terms of communication. The structure in our case is “socially coherent”, with a strong core integrating the diverse agents. In functional terms, such a structure can be interpreted as highly integrative, since each agent is easily involved in a communicative exchange with all other members in the organization.[8]

A second step in our analysis is concerned with observing the involvement of individual agents in the organizational structure. The overall structure is a result of the communication between individual agents, and role constructs serve as interfaces between the individual and the social layer. To empirically observe roles from communication, we once more made use of the network of utterances and replies in the discourse. Research in SNA has demonstrated the importance of the degree measure, i.e. the number of relations an agent has with other agents. We use the different types of relationships (one-way directed, mutual) to identify typical patterns in the network of communications. We count the proportion of in-degrees (i.e., the number of other agents referring to a particular agent), out-degrees (i.e., the number of others an agent refers to) and reciprocal ties for each individual agent. To find the most typical combination of relationships, we analyzed the results for the 228 agents by means of a hierarchical clustering algorithm (employing a “single linkage” method). The result of the clustering are 8 types of agent roles[9] that have a peculiar combination of in-degrees, out-degrees and reciprocal ties, representing a type of communicative behavior, i.e. a role in the organization.

The most frequent role we observed (besides the “passive” agents in the discourse) was that of “supporters”. “Supporters” in a discussion are agents that mostly refer to contributions by other agents, without invoking much response and without engaging in mutual exchanges. Similar to links in the World Wide Web [cf. 16], they support the original contributions by drawing more attention to them. They function as a sort of “resonance body” within the discourse.[10]

[8] The importance of such an integrative structure of communication is shown by Erickson. In an empirical study [9], she demonstrates that conversations about integrative topics such as, e.g., sports bridge the social gaps between organizational members, and enhance the integration of new members and the diffusion of information.

[9] See table 1 and figure 2. The results represent 97,5% of the information in the data.

[10] Rauch has analyzed a number of (face-to-face) meetings of large groups. He concludes that in all such communications, a small number of highly active central agents is backed up by a large number of more passive actors [27].

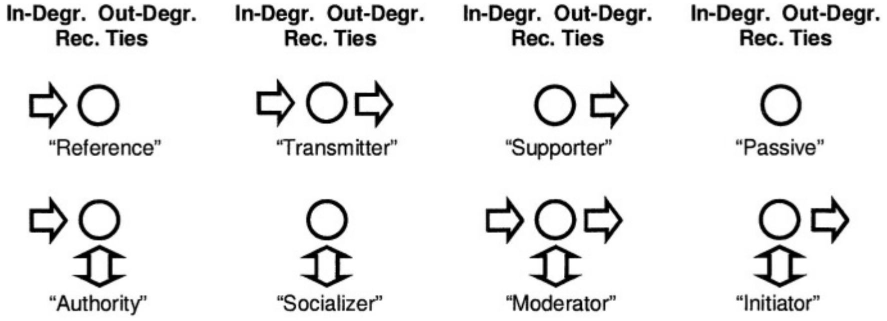


Fig. 2. Typical roles of communicative behavior. The arrows show for each role type the typical mixture of communicative relations: referrals received from other agents *(in-degrees),* referrals to other agents *(out-degrees),* and mutual exchanges *(reciprocal ties)*

The counterpart to these supporters is played by so-called "transmitters", "authorities" and "references", who all share a high proportion of in-degrees indicating that other agents refer to them frequently. "Transmitters" in this typology are agents that have relatively high in- and out-degrees, but do not engage in communication based on mutuality. They mainly pass on information. Their low level of engagement also shows up in the small amount of time they spent participating in the debate (8,3 days in average of the 42 days the debate lasted). "Authorities" and "references" both can be seen as highly visible and influential agents, in that their contributions are taken on by other agents and their ideas are communicated further on in the discourse. It is not surprising that all actively participating professors played one of these two roles – their high status within in the organization made many students reply to their contributions.

In contrast to those roles with high levels of in-degrees, "moderators" (in our case, this type empirically included the real moderator of the discourse), "socializers" and "initiators" have more mutual links. This indicates that besides referring to others or being referred to, they establish social ties to other agents that typically last for a number of communicative exchanges. "Moderators" have the most balanced combination of links. They play a very important role in the discourse, with a high level of engagement and with the largest number of contributions. Their high centrality score, a network measure of importance,[11] reflects this dominant position. "Socializers" and "initiators" show a different behavior. Both turn most of their relations into mutual exchanges. "Initiators" also strongly refer to third party agents. Typical "initiators" are students engaged in the students union, with a high visibility among the other students and with high engagement to stimulate the debate.

[11] We applied the so-called "betweenness-centrality" index [31].

Table 1. Properties of communicative role types

Role type	N	No. contr.	% In-links	% Out-links	% Rec. links	Centrality
"Passive"	124	0	-	-	-	0,0
"Supporter"	38	9,7	0,15	0,72	0,14	71,7
"Transmitter"	10	2,6	0,49	0,49	0,02	6,3
"Moderator"	3	79,7	0,41	0,15	0,44	1227,0
"Socializer"	6	4,2	0,17	0,17	0,66	34,5
"Initiator"	7	27,4	0,08	0,47	0,45	529,3
"Authority"	7	26,6	0,55	0,10	0,35	115,5
"Reference"	33	3,2	0,84	0,11	0,04	37,3
Overall	228	5,3	0,43	0,40	0,17	54,5

As explained above, each role serves to stabilize a type of behavior, and thus allows to build expectations about future reactions of the agents playing the role. Observing role models from communication with the help of typical combinations of relationships is an elegant way to gain such knowledge without having to make sophisticated assumptions about the agent's architecture. In an open multi-agent system for example, role models may provide the only available information about how other agents will react to an offer. These expectations enrich the information about the addresses an agent knows about, thereby turning information into knowledge which can guide decisions about with whom to communicate what.

Knowledge about organizational structures can be enriched by analyzing the relations between roles and the choice of communication channels. Organizational structures manifest themselves as stable channels along which communication typically proceeds. As part of the organizational structure, these channels exist not between individual persons, but between groups of persons with specific roles. Empirically, these channels can be observed by means of another SNA method called "blockmodelling". A blockmodel represents a reduced graph of the network, in which single persons are grouped into blocks according to the similarity of their communicative links. The relations between the blocks represent the structure of communication channels in the network, and each block contains all agents with structurally equivalent positions in the network.

The block model for our case study was calculated using the CONCOR algorithm [33]. After four iterations, the reduced graph consists of eight interconnected blocks and one isolated block. This isolated block contains all "passive" participants. The other blocks with the active participants in the discourse are more or less strongly connected to each other (see figure 3 and table 2 below). In order to interpret the results of the blockmodel analysis, we also computed the centrality indices for each block in the reduced graph.

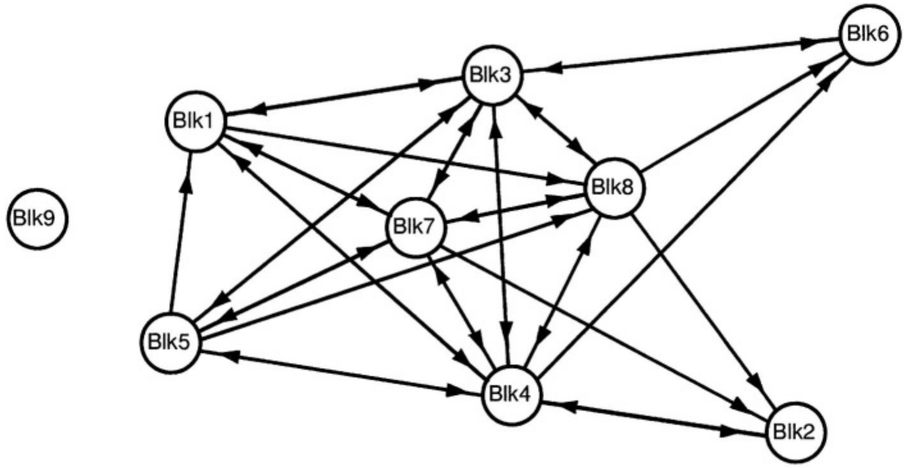


Fig. 3. Reduced graph of the communicative network, showing blocks of agents with structurally equivalent relations. E.g., all agents in block 9 are isolated in the flow of communication

Table 2. Blocks characterized by combination of role types (proportions are given as percentages) and centrality score (absolute figures)

Role Types	Blk 1	Blk 2	Blk 3	Blk 4	Blk 5	Blk 6	Blk 7	Blk 8	Blk 9	Overall
"Passive"	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	100,0	54,4
"Supporter"	54,5	42,9	21,7	38,9	77,8	0,0	38,1	20,0	0,0	16,7
"Transmitter"	27,3	14,3	4,3	0,0	0,0	0,0	19,0	10,0	0,0	4,4
"Moderator"	0,0	0,0	0,0	11,1	0,0	0,0	0,0	10,0	0,0	1,3
"Socializer"	0,0	0,0	8,7	5,6	0,0	20,0	9,5	0,0	0,0	2,6
"Initiator"	9,1	0,0	0,0	5,6	0,0	0,0	19,0	10,0	0,0	3,1
"Authority"	0,0	0,0	8,7	22,2	0,0	0,0	0,0	10,0	0,0	3,1
"Reference"	9,1	42,9	56,5	16,7	22,2	80,0	14,3	40,0	0,0	14,5
Block centrality	0,0	0,0	8,0	9,3	0,0	0,0	2,3	2,3	0,0	-

Blocks 3, 4, 7 and 8 form the core of the reduced graph. They maintain mutual relationships with each other and are connected to most of the other blocks. Block 3 and 4 are in an even more important position because they are the only ones in the network receiving referrals from block 2 or 6. The importance of the blocks in the reduced graph can be derived from the centrality scores. Block 3 and 4 score highest on the centrality index and can be called a "central core" (while block 7 and 8 are the "extended core"). They also contain the most active participants in the discourse (except one case located in block 7). The persons in these blocks have very heterogeneous communicative roles. Each block consists of a mixture of roles with a relatively high degree of "references" in block 3 and 8 and a relatively high degree of "supporters" in block 4 and 7.

Block 1 and 5 are less closely integrated in the communication channels since they are not mutually linked and completely lack a connection to blocks 2 and 6. However, they share reciprocal ties with most of the central blocks. Block 8 has a special position with respect to these two blocks: it does not reciprocate the links from block 1 and 5, and it is the only short connection of these blocks to the peripheral blocks 2 and 6. That means in terms of communication channels, block 8 is a sort of gatekeeper in the flow of information from the one periphery (block 1 and 5, which can be called "supporting periphery") to the other (block 2 and 6, which are the "activating periphery").[12] This interpretation is confirmed by the composition of the blocks: agents in block 1 and 5 typically have "supporter" roles, mainly referring to other agents without stimulating too many responses. Such responses come only from the agents in the central blocks, not from other more peripheral blocks.

Two other blocks are also only loosely connected to the center, yet in a different way: Block 2 and 6, comprising mostly "references" (and in block 2 also "supporters"), refer to the two most central blocks, but receive references only via block 8. That means, part of the flow of communication in the discourse starts in the center (block 3 and 4), the "activating periphery" takes up the thread, and block 8 acts as a sort of catalyst and distributor, opening the discussion by stimulating responses from the rest of the active participants, the central blocks as well as the "passive periphery".

As an overall picture, the reduced graph of the network of communication shows that the roles identified previously are related to specific channels of communication, and together form an important part of the organizational structure. While "supporters" are mainly found in the periphery, especially in block 1 and 5, "authorities" as well as "references" are found in the center, especially in block 3 and 4. The other blocks in the model are more heterogeneous with a balanced mix of different role types.

By comparing the role analysis with the positions of the blockmodel, an observing agent can build communicative strategies for retrieving or spreading information. I.e., in order to be well informed about specific issues, an agent does not need to get in touch with the few most involved users. Instead, he can use the large number of "supporters". Also, to spread information into the communicative network, "socializers" and "initiators" with central position, e.g., in block 7, are the right starting point. Thus, we can use the model of communication channels to communicate more effectively with respect to the overall social context.

[12] The importance of such gatekeeping agents is stressed by Burt's analysis of "structural holes" [5].

5 Summary and Conclusions

The results of this empirical study illustrate how the paradigmatic shift in social theory from interaction to communication can be instructive for MAS design. We first clarified basic concepts of organizational structure, like persons, roles and communication channels, on a theoretical basis. From our systems theoretical approach, such organizational structures are regarded as stable communicational patterns, emerging from highly dynamic behavior. If such structures can be observed, the knowledge generated from this observation can serve MAS designers to design flexible and adaptive agents for large-scale open systems. In our case study of intra-organizational communication, we demonstrated how our approach can be used to observe the social structures of organizations. We employed different measures from SNA to analyze three important organizational phenomena:

- The overall structure of communication in an organization was visualized by means of component analysis.
- Typical roles of communicative behavior were identified by analyzing the pattern of the agents' relationships with other agents.
- Communication channels between groups of agents were observed with the help of the blockmodelling technique and an analysis of the correlation between communication channels and role types.

These results serve primarily as illustrations. Nevertheless, they demonstrate the potential of systems theory in combination with SNA methods to give a very clear picture of emergent structures in social systems like, e.g., organizations. Given the case that – as in open, large-scale MAS – agents have to cope with highly dynamic, fluctuating environments, we think that the observation of such patterns as we have shown in our analysis is of high importance for an agent's ability to operate socially and intelligently. Observations like the ones made in our example can help agents to grasp the basic structures of the system they are operating in, to adapt flexibly to a changing environment, and to model their social exchanges in an efficient and reliable way.

Our sociological starting point is that Luhmann's "Theory of Social Systems" offers a new perspective on organizations and allows us to analyze and interpret communicative data in a novel way. This theory represents a framework that fits very well with the methodological approach of social network analysis. Both can be used to further improve the design methodology of MAS. Roles and positions in existing MAS could be monitored and analyzed by a sort of "social mirror" to help new incoming agents to develop an efficient communication strategy, or – in terms of social system theory – to increase the probability of the unlikely process of the stabilization of communicational patterns.

As an outlook towards the practical implications of our approach, software agents working as assistants in organizations could look for adequate contact partners or they could summarize past communication processes. For the organization, this would mean to build a kind of socially enhanced organizational memory, in which agents co-act with humans to establish and permanently redesign organizational structures. The management of such an organization could use the information provided by such observers as a barometer of the intra-organizational opinion. Last but not least, such agents could help to manage "knowledge communities" [8]. These communities are informal networks within organizations that have to be handled carefully, without interventions from top-down, but rather with a bottom-up approach to governance.

References

1. Ackerman, M.S., Halverson, C.: Organizational Memory: Processes, Boundary Objects, and Trajectories. Proceedings of the 32nd Hawaii International Conference on System Sciences, Maui, Hawaii, January 5–8, IEEE Computer Society, Los Alamitos et al., (1999)
2. Albrecht, S.: Structuring Large-Scale Online Debates – Making Use of Network Analysis Methods. Paper presented at the Sunbelt XXI International Social Network Conference, Budapest, Hungary, April 25–28 (2001)
3. Bond, A.H., Gasser, L.: An Analysis of Problems and Research in DAI. In: Bond, A.H., Gasser, L. (eds.): Readings in Distributed Artificial Intelligence. Morgan Kaufman, San Mateo, CA (1988) 3–35
4. Brauer, W., Nickles, M., Rovatsos, M., Weiß, G., Lorentzen, K.F.: Expectation-Oriented Analysis and Design. In: Wooldridge, M.J., Weiß, G., Ciancarini, P. (eds.): Agent-Oriented Software Engineering II. Second International Workshop, AOSE 2001, Montreal, Canada, May 29, Lecture Notes in Computer Science, Vol. 2222. Springer-Verlag, Berlin Heidelberg New York (2002) 226–234
5. Burt, R.S.: Structural Holes: The social structure of competition. Harvard University Press, Cambridge, MA (1992)
6. Carley, K.M.: On the Evolution of Social and Organizational Networks. In: Andrews, S.B., Knoke, D. (eds.): Networks In and Around Organizations. JAI Press, Stamford (1999) 3–30
7. Carley, K.M., Gasser, L.: Computational Organizational Theory. In: Weiss, G. (ed.): Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence. MIT Press, Cambridge, MA, London (1999) 299–330
8. Diemers, D.: Virtual Knowledge Communities. Erfolgreicher Umgang mit Wissen im digitalen Zeitalter. Difo-Druck, Bamberg (2001)
9. Erickson, B.H.: Culture, Class, and Connections. American Journal of Sociology 102 (1996) 217–251
10. Ferber, J.: Multi-agent Systems. An Introduction to Distributed Artificial Intelligence. Addison-Wesley, Harlow (1999)
11. Fox, M.: An Organizational View of Distributed Systems. IEEE Transactions on Systems, Man and Cybernetics 11 (1981) 70–80
12. Gasser, L.: Social Conceptions of Knowledge and Action: DAI Foundations and Open Systems Semantics. Artificial Intelligence 47 (1991) 107–138
13. Hewitt, C.E.: Open Information Systems Semantics for Distributed Artificial Intelligence. Artificial Intelligence 47 (1991) 79–106
14. Kiesler, S., Sproull, L.: Connections. New Ways of Working in the Networked Organization. MIT Press, Cambridge MA, London (1991)
15. Kirn, S., Gasser, L.: Organizational Approaches to Coordination in Multi-Agent Systems. TU Ilmenau, Wirtschaftsinformatik 2, Working Paper No. 9, March (1998)
16. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46 (1999) 604–632

17. Lorentzen, K.F., Nickles, M.: Ordnung aus Chaos – Prolegomena zu einer Luhmann'schen Modellierung deentropisierender Strukturbildung in Multiagentensystemen. In: Kron, T., Junge, K., Papendick, S. (eds.): Luhmann modelliert. Ansätze zur Simulation von Kommunikationssystemen. Leske + Budrich, Opladen (2002) 55–113
18. Luehrs, R., Malsch, T., Voss, K.: Internet, Discourses and Democracy. In: Terano, T. et al. (eds.): New Frontiers in Artificial Intelligence. Joint JSAI 2001 Workshop Post-Proceedings. Lecture Notes in Computer Science, Vol. 2253. Springer-Verlag, Berlin Heidelberg New York (2001) 67–74
19. Luehrs, R., Albrecht, S., Lübcke, M., Hohberg, B.: How to Grow? Online Consultation about Growth in the City of Hamburg: Methods, Techniques, Success Factors. To appear in: Proc. of the 2nd Int. Conference on Electronic Government (EGOV 2003), September 1–5, Prague, Czech Republic. Springer-Verlag, Berlin Heidelberg New York (2003)
20. Luhmann, N.: Social Systems. Stanford University Press, Stanford, CA (1995)
21. Luhmann, N.: Die Gesellschaft der Gesellschaft. Suhrkamp, Frankfurt/M. (1997)
22. Luhmann, N.: Organisation und Entscheidung. Westdeutscher Verlag, Opladen (2000)
23. Malsch, T.: Naming the Unnamable: Socionics or the Sociological Turn of/to Distributed Artificial Intelligence. Autonomous Agents and Multi-Agent Systems 4 (2001) 155–186
24. March, J.: Footnotes to Organizational Change. Administrative Science Quarterly 26 (1981) 563–577
25. March, J.: Exploration and Exploitation in Organizational Learning. Organization Science 2(1991)71–87
26. Nohria, N., Eccles, R.G. (eds.): Networks and Organizations. Structure, Form and Action. Harvard Business School Press, Boston, MA (1992)
27. Rauch, H.: Partizipation und Leistung in Großgruppen-Sitzungen. In: Neidhardt, F. (ed.): Gruppensoziologie. Perspektiven und Materialien. Westdeutscher Verlag, Opladen (1983), 256–274
28. Rogers, E.M., Agarwala-Rogers, R.: Communication in Organizations. Free Press, New York (1976)
29. Sabater, J., Sierra, C: Reputation and Social Network Analysis in Multi-Agent Systems. In: Proc. of the the First International Joint Conference on Autonomous Agents and Multiagent systems (AAMAS'02), Bologna, Italy, July 15–19 (2002) 475–482
30. Strübing, J.: Bridging the Gap. On the Collaboration between Symbolic Interactionism and Distributed Artificial Intelligence in the Field of Multi-Agent Systems Research. Symbolic Interaction 21 (1998) 441–463
31. Wasserman, S., Faust, K.: Social Network Analysis. Methods and Applications. Cambridge University Press, Cambridge et al. (1994)
32. Wellman, B.: Computer Networks as Social Networks. Science 293 (2001) 2031–2034
33. White, H.C., Boorman, S.A., Breiger, R.L.: Social Structures from Multiple Networks. I. Blockmodels of Roles and Positions. American Journal of Sociology 81 (1976) 730–779
34. Wooldridge, M.: Intelligent Agents. In: Weiss, G. (ed.): Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence. MIT Press, Cambridge, MA, London (1999) 27–77
35. Yu, B., Singh, M.P.: Search in Referral Networks. Paper presented at the Int. Workshop on Regulated Agent-Based Social Systems, RASTA'02, Bologna, Italy, July 15–19 (2002) (published in this volume).
36. Zambonelli, F., Jennings, N.R., Wooldridge, M.: Organisational Rules as an Abstraction for the Analysis and Design of Multi-agent Systems. International Journal of Software Engineering and Knowledge Engineering 11 (2001) 303–328

On How to Conduct Experimental Research with Self-Motivated Agents

Luis Antunes and Helder Coelho

Faculdade de Ciências, Universidade de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal
{xarax, hcoelho}@di.fc.ul.pt

Abstract. We argue that experimental methodologies are harder to apply when self-motivated agents are involved, especially when the issue of choice gains its due relevance in their model. We use a choice-oriented agent architecture to illustrate a means of bridging the distance between the observer and the actors of an experiment. Traditional experimentation has to give way to exploratory simulation, to bring insights into the design issues, not only of the agents, but of the experiment as well. The role of its designer cannot be ignored, at the risk of achieving only obvious, predictable conclusions. We propose to bring the designer *into* the experiment. To accomplish that, we provide a value-based model of choice to represent the preferences of both entities. This model includes mechanisms that allow for explicit bonds between observer and observed. We use the findings of extensive experimentation with this model to compare current experimental methodologies in what concerns evaluation itself.

1 Context

> "Artificial intelligence [is] the problem of designing agents that *do the right thing.*" [23, page 2, original italics]

A new scientific synthesis has been taking form under the name of artificial intelligence [27]. This young discipline has recently started to rearrange itself under the keynote concept of agent [22]. Agents can be seen as unwanting actors, but gain additional technological interest and use when they have their own motivations, and are left for autonomous labour [7]. Of course, ethical considerations about the role of their designers are required, to constrain the design space of their architectures. But norms may be grounded on reliable experimentation. It is not only a question of metrics (preferences, ranking functions), but the whole methodology is a key issue.

In informatics, no-one is completely assured that a program does the "right thing," or all faulty behaviours are absent. In agent technologies, we would like to discover when a creature may turn into a mad (paranoic) one, i. e., know how to switch off inconvenient performance and tune its behaviour in deep detail. If agents are to be used by someone, *trust* is *the* key issue. But, how can we trust a agent that pursues its own agenda to accomplish some goals of ours [7]?

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 31–47, 2004.

Autonomy deals with the agents' freedom of choice, and choice leads to the agents' behaviour through two specific phases in the decision process. Unlike BDI (beliefs-desires-intentions, cf. [3]) models, where the stress is given on the technical issues dealing with the agents pro-attitudes (what can be achieved, how can it be done), in the BVG (beliefs-values-goals, cf. [3]) model, the emphasis is given on choice machinery. Choice is about *which* goals to pursue (or, where do the goals come from), and how the agent *prefers* to pursue them (or, which options the agent wants to pick).

In [1,3,2], we have defended a model of choice that depends on the idea of multi-dimensional evaluation of a choice situation to accomplish an enhanced adaptivity to a dynamic and complex environment. These dimensions (which we have called *values*), are used to select which goals to pursue (through the information of mechanisms for goal adoption), and also which sub-goals are to be preferred and selected for execution.

The central question is evaluation of the quality of decision. If the agent aims at optimising this measure (which can in turn be multi-dimensional), why does s/he not use it for the decision in the first place? And, should this measure be unidimensional, does it amount to a utility function (which would configure the "totilitarian" view: maximising utility as the *sole* motivation of the agent)?

This view, however discredited since the times of the foundation of artificial intelligence (which was founded against an utilitarian view of rationality [24]), still prevails in many approaches, even through economics or the social sciences (cf. [6,13]).

But in artificial intelligence, the issue of methodology is still in order. Other sciences, even older ones, are still looking for their identity, and to that end, the very concept of computer seems to contribute: it changes the notion of what can be done (calculate), changes the (workable) object of the science, changes the methods. The complexity revolution has only begun, with the gradual and (more and more) systematic use of modern calculation means.

The fact that artificial intelligence depends on this new tool (the computer) does not ease up this transition. If artificial intelligence possessed from the very beginning the means that are believed to be able to make (generate) the difference, it takes time to find the course, and the methods. The inspiration from social sciences that is on the origin of multi-agent systems is an example of an import. Meanwhile, those same systems are already being exported to other sciences (as economics and sociology), and precisely as methodological aids.

In the next section, we summarise our choice framework, and state the problem of evaluating the results of the agents' decisions. In section 3, we compare two methodologies for conducting experiments for multi-agent systems. We conclude that the issue is not conpletely solved by either one, and note the similarities between evaluation of the results by the designer, and adaptation by the agents. In section 4 we propose two answers for the issue of assessing experimental results. The combination of both approaches, bringing the designer's insights and conjectures into the setting of experiments, fits well into the notion of pursuing exploratory simulation. In the last two sections, we briefly present our experimental results, and finally conclude by exalting the advantages of explicitly connecting the experimenter's and the agents' evaluative dimensions.

2 Choice and Evaluation

The role of value as a new mental attitude towards decision is twofold. On the one hand, values provide a reference framework to represent agent's preference during deliberation (the pondering of options candidate to contribute to a selected goal). On the other, values help inform choice, the final phase of decision, when the agent has to pick an option from the ordered set of options provided by the deliberation phase. To this aim, a probability distribution can be defined by using the relevant values for the situation.

In the BVG choice framework, the agent's system of values evolves as a consequence of the agent's assessment of the results of previous decisions. Decisions are evaluated against certain dimensions (that could be the same previously used for the decision or not), and this assessment is fed back into the agent's mind, by adapting the mechanisms associated with choice, especially the ones related to values. This is another point that escapes the traditional utilitarian view, where the world (and so the agent) is static and known. BVG agents can adapt to an environment where everything changes, including the agent's own preferences (for instance as a result of interactions). This is especially important in a multi-agent environment, since the agents are autonomous, and so potentially sources of change and novelty.

The evaluation of the results of our evaluations becomes a central issue, and this question directly points to the difficulties in assessing the results of experiments. We would need meta-values to evaluate those results, but that calls for a designer, and amounts to looking for emergent phenomena. But if those "higher values" exist (and so they are the important ones) why not use them for decision? This dilemma clearly shows the *ad hoc* character of most solutions, and it is difficult to escape it.

We can conceive two ways out. The first is the development of an ontology of values, to be used in some class of situations as qualitative markers (norms). Higher or lower, values have their place in this ontology, and their relations are clearly defined. For a given problem the relevant values can be identified and used, and appropriate experimental predictions postulated and tested.

When tackling the issue of choice, the formulation of hypotheses and experimental predictions becomes delicate. If the designer tells the agent how to choose, how can he not know exactly how the agent will choose? To formulate experimental predictions and then evaluate to what extent they are fulfilled becomes in this case a spurious game: it amounts to perform calculations about knowledge and reasons, and not to judge to what extent those reasons are the best reasons, and correctly generate the choices. We return to technical reasons for behaviour, in detriment of the will and the preferences of the agent.

Consequently, the second solution is subtler. By situating the agent in an environment with other agents, autonomy becomes a key ingredient, to be used with care and balance. The duality of value sets becomes a necessity, as agents cannot access values at the macro level, made judiciously coincide with the designer values. The answer is the designer, and the problem is methodological. The BVG update mechanism provides a way to put to test this liaison between agent and designer.

The designer's model of choice cannot be the model of perfect choice against which the whole world is to be evaluated. It is our strong conviction that the perfect choice does not exist, because characters perfectly embody a specific set of physical and personality traits. All depends on the adequate fiction (script). It is a model of choice to be compared to another (human) one playing an identical role, by using criteria that in turn may not be perfect.

3 Experimental Methodologies

When Herbert Simon received his Turing award, back in 1973, he felt the need to postulate "artificial intelligence is an empirical science." The duality science/engineering was always a mark of artificial intelligence, so that claim is neither empty nor innocent. Since that time, there has been an ever-increasing effort in artificial intelligence and computer science to experimentally validate the proclaimed results. The interdisciplinary site of artificial intelligence was not always equally prone to imports of scientific methods from other disciplines. So, theoretical demonstration was for decades more used than empirical one. And exemplification (*one-shot experiment*) more common than demonstration.

[11], followed by [21], try to define the general lines of an experimental method for artificial intelligence. Controlled experimentation aims at solidifying the scientific discipline, through the variation of the features of a system or its environment, for posterior measure of the effect of those variations in the performance of the system. The worry of [21] is that experimentation led by testbeds and benchmarks provides only a comfortable illusion of scientific progress, but not significant and generalisable results. Comparative measures are valuable for certain ends, but only constitute scientific progress if they suggest or provide evidence for theories that can explain the differences in performance.

Steve Hanks and Martha Pollack debated over the question of realism and result generalisation. Any experimentable phenomenon has as basis a model of the real one. But simplifications necessary to the modelling process can be so strong that the resulting model is very far from the original model. Such an irrealistic model, despite allowing for controlled experimentation, is almost useless, since it will hardly allow the generalisation of the results to real world systems embedded in complex environments.

To achieve this balance between simplicity and realism, Hanks proposed to focus on more realistic systems and environments, and to conduct experiments directly over them. The danger of experimentation on overly simplistic models is to turn attention to the *experimental process* itself, instead of the *ideas* that are supposed to be tested. That is, the danger is the seduction of solving problems one can understand, instead of problems that are interesting.

From another standpoint, Pollack suggested it is enough to keep systematicity in experiments, and look for inspiration on how to generalise results in other sciences, with a greater experimentation tradition. And she argued that experiments, even if simple (and simplicity permits experimental control), can suggest additional ones, that is, experimentation is an iterative process, and a part of the experimental process is precisely to refine the mapping between a theory and its realisation in implemented systems.

A system, however big and realistic, is always a model, and so there will always be a distance separating it from reality. The third author, Paul Cohen, ended up giving the methodological answer to the problem of generalising the results of experiments. Acknowledging that empirical results are seldom general, Cohen insisted that nothing prevents the researcher from "inventing general theories as interpretations of results of studies in simulation testbeds, and nothing prevents [him] (...) from designing additional studies to test predictions of these theories in several simulation testbeds" [21, page 39].

3.1 A Methodology for Principled Experimentation

Simulation testbeds (and so controlled small scale experimentation) have a relevant role in three phases of research. In an exploratory phase, to provide the environment where the agents will be inserted; in a confirmation phase, by more strictly defining the characterisations of behaviours, and testing specific hypotheses; in a generalisation phase, by trying to replicate the results.

MAD (Modelling, Analysis and Design) involves seven activities [11]: (1) evaluate the environmental factors that affect behaviour; (2) model the causal relations between system design, its environment, and its behaviour; (3) design or redesign a system (or part of one); (4) predict how the system will behave; (5) run experiments to test predictions; (6) explain unexpected results and modify the models and design of the system; and (7) generalise models to classes of systems, environments and behaviours.

In [12], Cohen sustains that the goals of the study can be exploratory, test scientific hypotheses, provide calibration data, adequate model parameters, etc. Demonstrations (traditional in artificial intelligence) show only how something can be made to work, are not necessarily exploratory, or test hypotheses, or estimate parameters. The exploratory dimension follows Cohen idea, of defining designs around ideas, instead of valuing premature experimentation.

Cohen states the fundamental question to link this methodology to the concept of experiment with self-motivated agents we envisage. The third criterion to evaluate the design of experiments is: "What are the criteria of good performance? Who defines these criteria?"

The answer to these questions is an invitation to consider rationality itself, and its criteria. The fact that rationality is situated (in some sort of fiction) most times imposes the adoption of *ad hoc* decision criteria. But the evaluation of the results of experiments is not intrinsically different from the evaluation the agents conduct of their own performance (and upon which they base their adaptation). In particular, there was always a designer defining both types of evaluation.

So the question comes natural: why would the design of sonic component be "better" than the other (and support one "right thing")? Most times there is no reason at all, and the designer uses the same criteria (the same "rationality") either for the agent's adaptation or for the evaluation of its performance. If a better way of choosing is found, both components are redesigned, but always together. The alternative to this scenario amounts to look for and study emergent properties of the systems, and we will tackle it in the following sections. As to MAD methodology, it was critically reviewed in [10] and the application of the resulting methodology (extended MAD) will be reassessed in section 5.

3.2 A Methodology from the Social Sciences

The problem we have just posed about the duality experiment evaluation *versus* evaluation by the agents themselves, is not different from the above mentioned problem of the generalisation of the results of experiments: to what extent isn't our evaluation of experiments more centred in the design of the agents that participate in it?

When both these questions seem alike, and the experiments criteria seem to be themselves object of the experiment, perhaps we have reached the limits of simplification, and reductionist techniques. An alternative is to revert positions: exchange reductionist analyses for holistic ones (integrating syntheses), and simplification for complexity [27].

In multi-agent systems, the greatest development of this tendency happened in the interaction with the social sciences, and had the greatest reach with the opposition of simulation to controlled experimentation. Multi-agent systems get their inspiration in eminently social phenomena. The first metaphor to try out is the social one, motivated by the way (mainly) human agents organise themselves. But artificial intelligence brings the advantage of computational models, hence easily manipulable. So the second metaphor bases itself on the first one, but introduces variation.

Quickly, social scientists understood the potential of the return of this inspiration export. The notion of agent and computational simulation are the master beams of the new complexity science [14].

Computational simulation is methodologically appropriate when a social phenomenon is not directly accessible [19]. One of the reasons for this unaccessibility is the target phenomenon being so complex that the researcher cannot grasp its relevant elements. Simulation is based in a more observable phenomenon than the target one. Often the study of the model is as interesting as the study of the phenomenon itself, and the model becomes a legitimate object of research [13]. There is a shift from the focus of research of natural societies (the behaviour of a society model can be observed "in vitro" to test the underlying theory) to the artificial societies themselves (study of possible societies). The questions to be answered cease to be "what happened?" and "what may have happened?" and become "what are the necessary conditions for a given result to be obtained?," and cease to have a purely descriptive character to acquire a prescriptive one. A new methodology can be synthesised, and designated "exploratory simulation"

[13]. The prescriptive character (exploration) cannot be simplistically resumed to a optimisation, such as the descriptive character is not a simple reproduction of the real social phenomena.

In social sciences, an appropriate methodology for computational simulation could be the one outlined by Nigel Gilbert [18]: (1) identify a "puzzle," a *question* whose answer is unknown; (2) *definition* of the target of modelling; (3) normally, some *observations* of the target are necessary, to provide the parameters and initial conditions of the model; (4) after developing the model (probably in the form of a computer program), the *simulation* is executed, and its results are registered; (5) *verification* assures the model is correctly developed; (6) *validation* ensures that the behaviour of the model corresponds to the behaviour of the target; and (7) finally, the *sensitivity analysis* tells how sensitive the model is to small changes in the parameters and initial conditions.

We are not far from MAD methodology, but there are fundamental differences: in MAD there is no return to the original phenomenon. The emphasis is still on the system, and the confrontation of the model with reality is done once and for all, and represented by causal relations. All the validation is done at the level of the model, and the journey back to reality is done already in generalisation. In some way, that difference is acceptable, since the object of the disciplines is also different. But it is Cohen himself who asks for more realism in experimentation, and his methodology fails in that involvement with reality.

But, is it possible to do better? Is the validation step in Gilbert's methodology a realist one? Or can we only compare models with other models and never with reality? If our computational model produces results that are adequate to what is known about the real phenomenon, can we say that our model is validated, or does that depend on the source of knowledge about that phenomenon? Isn't that knowledge obtained also from models? For instance, from results of questionnaires filled by a representative sample of the population – where is the real phenomenon here? Which of the models is then the correct one?

The answer could be in [25]: social sciences have an exploratory purpose, but also a predictive and even prescriptive one. Before we conduct simulations that allow predictions and prescriptions, it is necessary to understand the phenomena, and for that one uses exploratory simulation, the exploration of simulated (small) worlds. But when we do prediction, the real world gives the answer about the validity of the model.

Once collected the results of simulations, they have to be confronted with the phenomenon, for validation. But this confrontation is no more than analysis. With the model of the phenomenon to address and the model of the data to collect, we have again a simplification of the problem, and the question of interpretation returns, which we have already found in localised experimentation. It certainly isn't possible to suppress the role of the researcher, ultimate interpreter of all experiments, be it classical or simulation. The bottom-up approach which forms the basis of computational simulation forces us to consider the concept of emergence.

When conducting experiments and simulations, it is a constant worry of the designer to verify if the so-called emergent behaviours wouldn't be pre-programmed, in the sense of being an inevitable consequence of the way the agents were built. Gilbert [17] provides a criterion to distinguish emergent behaviour (in Gestalt sense, according to Castelfranchi's account [8]) from behaviour predictable from the individual characteristics of the agents: it should not be possible to analytically derive the global emergent behaviour solely from the consideration of the agents' properties. That is, the analysis *has* to fail, and the simulation be inevitable to discover those properties. But emergence may not even be a stable or interesting property of systems: what is interesting are the system macro-properties and their relations with its micro-properties [17].

We can redraw the problem we have described above (evaluation of the results by the designer *versus* adaptation by the agents) in the more restrained panorama of Castelfranchi's different emergencies. The observer of the agent's performance becomes the agent itself, and representational-emergence is confused with Gestalt-emergence.

We will introduce this exact relation between the designer of the experiment and its participant (agent). The information the agent uses to update its choice machinery may or not be related to the information the designer is interested in observing. And the performance measures of agents and societies may or not, correspondingly, be related to the evaluation measures of experiments. But remember the discussion in the beginning of this paper, neither ones are the *perfect* measures.

4 Two Answers

In this section we will present two different answers for the problem of analysing (and afterwards, generalising) the results of the experimentation, which we have already argued to have quite a strong connection to the problem of *improving* the agents performance as a result of evaluation of the previous choices.

The explicit consideration of the relevant evaluative dimensions in decision situations can arguably provide a bridge between the agent's and the experiments designer's mind. In a model such as BVG [3], the agent's choice mechanisms are fed back with a set of multi-dimensional update values. These dimensions may or not be the same that were used to make the decision in the first place. If these dimensions should be different, we can identify the ones that were used for decision with the interests of the agent, and the ones used for update with the interests of the designer. And moreover, we have an explicit link between the two sets of interests. So, the designer is no longer left for purely subjective guessing of what might be happening, confronted with the infinite regress of ever more challenging choices. S/he can explore the liaisons provided by this choice framework, and experiment with different sets of preferences (desired results), both of hers and of the agents.

However, the problem remains of finally having to have a say about what really happened. In fact, both authors of this text have slightly different positions, and in the remainder of this section we will diverge to elaborate on those positions. Or, you can look at the following two subsections as different scientific (or better, scientifically philosophic) hypotheses about evaluation of evaluations.

4.1 Positivism: Means-Ends Analysis in a Layered Mind

We can postulate a positivist (optimistic) position by basing our ultimate evaluations on a pre-conceived ontology of such deemed relevant dimensions (or values). Having those as a top-level reference, the designer's efforts can concentrate on the appropriate models, techniques and mechanisms to achieve the best possible performance as measured along those dimensions.

It seems that all that remains is then optimisation along the desired dimensions, but even in that restrained view we have to acknowledge that it does not mean that all problems are now solved. Chess is a domain where information is perfect and the number of possibilities is limited, and even so it was not (will it ever be?) solved.

Alternatively, the designer can be interested in evaluating how the agents perform in the absence of the knowledge of what dimensions are to be optimised. In this case, several models can be used, and the links to the designer's mind can still be expressed in the terms described above.

The key idea is to approximate the states that the agent wishes to achieve to those that it believes are currently valid. This amounts to performing a complex form of means-ends analysis, in which the agent's sociality is an issue, but necessarily one in which the agent does not have any perception about the meta-values involved. Because that would reinstate the infinite regression problem.

The external evaluation problem can be represented in terms as complex as the experiment designer thinks appropriate. Suppose the scenario of a supermarket where types of behaviours are expected. How can we force our agent to acquire those behaviours? First, tuning the agent to move from some initial (IB) to a final behaviour (FB), due to reflexive markers (a sort of norms [16]). The tuning is controlled by an operator able to reduce the difference between IB and FB. In a BDI-like logical approach, evaluation can be as simple as answering the question "were the desired states achieved or not?," or as complicated as the designer desires and the decision framework allows to represent.

The choice mechanisms update becomes an important issue, for they are trusted to generate the desired approximation between the agent's performance (in whichever terms) and the desired one.

Interesting new architectural features recently introduced by Castelfranchi [8] can come to the aid of the task of unveiling these ultimate aims that justify behaviour. Castelfranchi acknowledges a problem for the theory of cognitive agents: "how *to reconcile the 'external' teleology of behamour with the 'internal' teleology governing it;* how to reconcile intentionality, deliberation, and planning with playing social functions and contributing to the social order." [8, page 6, original italics].

He then notes that "self-organising social processes – not being chosen – are indifferent, in principle, to the agents' or groups' goals and welfare; they are not necessarily (...) advantageous for something and somebody. Since the effects reproducing the behaviour are not realised and appreciated by the subject, there is no reason for assuming that they will necessarily be 'good' for his/her needs or aims, or good for society's aims." [8, page 35].

Following van Parijs [26], Castelfranchi defends *reinforcement* as a kind of internal natural selection, the selection of an item (e.g. a habit) directly within the entity, through the operation of some internal choice criterion.

And so, to meet our own idea of means-ends analysis, Castelfranchi proposes the notion of *learning,* in particular, reinforcement learning in cognitive, deliberative agents. This could be realised in a hybrid layered architecture, but not one where reactive behaviours compete against a declarative component. The idea is to have "a number of low-level (automatic, reactive, merely associative) mechanisms operate *upon* the layer of high cognitive representations" [8, page 22, original italics].

Damasio's [15] somatic markers, and consequent mental reactions of attraction or repulsion, serve to constrain high level explicit mental representations, as our reflexive markers do. This mental architecture can do without the necessity of an infinite recursion of meta-levels, goals and meta-goals, decisions about preferences and decisions. In this meta-level layer there could be no explicit goals, but only simple procedures, functionally teleological automatisms.

In the context of our ontology of values, the notion of attraction/repulse could correspond to the top level of the hierarchy, that is, the ultimate value to satisfy. Optimisation of some function, manipulation and elaboration of symbolic representations (such as goals), pre-programmed (functional) reactivity to stimuli, are three faces of the same notion of ending up the regress of motivations (and so of evaluations over experiments). This regress of abstract motivations can only be stopped by grounding the ultimate reason for choice in concrete concepts, coming from *embodied* minds.

4.2 Relativism: Extended MAD, Exploratory Simulation

As we explained, there are some problems in the application of MAD methodology to decision situations. MAD is heavily based on hypotheses formulation and predictions about systems behaviour, and posterior confrontation with experimental observations. We have already suggested that an alternative could be conjectures-led exploratory simulation.

The issues raised by the application of MAD deal with meta-evaluation of behaviours (and so, of underlying models). We have proposed an extension to MAD that concerns correction between the diverse levels of specification (from informal descriptions to implemented systems, passing by intermediate levels of more or less formal specification) [10]. This extension is based on the realisation of the double role of the observer of a situation (which we could translate here into the role of the agent and that of the designer).

The central point is to evaluate the results of agent's decisions. Since the agent is autonomous and has its own reasons for behaviour, how can the designer dispute its choices? A possible answer is that the designer is not interested in allowing the agent to use the best set of reasons. In this case what is being tested is not the agent, but what the designer thinks are the best reasons. The choice model to be tested is not the one of the agent, and the consequences may be dramatic in open societies.

In BVG, the feedback of such evaluative information can be explicitly used to alter the agents choice model, but also to model the mind of the designer. So, agents and designer can share the same terms in which the preferences can be expressed, and this eases up validation. The model of choice is not the perfect reference against which the world must be evaluated (we have already sustained that such a model does not exist), but just a model to be compared to another one, by using criteria that again might not be perfect.

This seems to amount to an infinite regress. If we provide a choice model of some designer, it is surely possible to replicate it in the choice model of an agent, given enough liberty degrees to allow the update mechanisms to act. But what does that tell us? Nothing we couldn't predict from the first instant, since it would suffice that the designer's model would be used *in* the agent. In truth, to establish a realist experiment, the designer's choice model would itself be subject to continuous evolution to represent his/her choices (since it is immersed in a complex dynamical world). And the agent's model, with its update mechanisms, would be "following" the other, as well as it could. But then,what about the designer's model, what does it evolve to follow? Which other choice model can this model be emulating, and how can it be represented?

Evaluation is harder for choice, for a number of reasons: choice is always situated and individual, and it is not prone to generalisations; it is not possible to establish criteria to compare choices that do not challenge the choice criteria themselves; the adaptation of the choice mechanisms to an evaluation criteria appears not as a test to its adaptation capabilities, but rather as a direct confrontation of the choices.

Who should tell if our choices are good or not, based on which criteria can s/he do it, why would we accept those criteria, and if we accept them and start making choices by them, how can we evaluate them afterwards? By transposing this argument to experimental methodology, we see the difficulty in its application, for the decisive step is compromised by this opposition between triviality (when we use the same criteria to choose and to evaluate choices) and infinite and inevitable regression (that we have just described).

Despite all this, the agent *cannot* be impotent, prevented from improving its choices. Certainly, human agents are not, since they keep choosing better (but not every time), learn from their mistakes, have better and better performances, not only in terms of some external opinion, but also according to their own. Not always the change in choices results from a tentative at improving, the agent may have only changed opinion. If one considers that the choice *made* by the agent is the best choice s/he could have done (in some sense, the *perfect choice* whose existence we refuse), changing opinion is the *only* way to improve.

As a step forward, and out of this uncomfortable situation, we can also consider that the agent has two different rationalities, one for choice, another for its

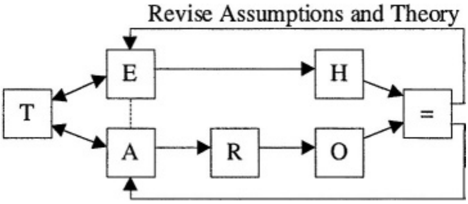


Fig. 1. Construction of theories. An existing theory (T) is translated in a set of assumptions (A) represented by a program and an explanation (E) that expresses the theory in terms of the program. The generation of hypotheses (H) from (E) and the comparison with observations (O) of runs (R) of the program allows both (A) and (E) to be revised. If finally (H) and (O) correspond, then (A), (E) and (H) can be fed back into a new revised theory (T) that can be applied to a real target (from [19]).

evaluation and subsequent improvement. One possible reason for such a design could be the complexity of the improvement function be so demanding that its use for common choices would not be justified.

To inform this choice evaluation function, we can envisage three candidates: (i) a higher value, or some specialist's opinion, be it (ii) some individual, or (iii) some aggregate, representing a prototype or group.

The first, we have already described in detail in the previous subsection: some higher value, at a top position in a ontological hierarchy of value. In a context of social games of life and death, survival could be a good candidate for such a value. As would some more abstract dimension of goodness or righteousness of a decision. That is, the unjustifiable (or irreducible) sensation that, all added up, the right (good, just) option is evident to the decider, even if all calculations show otherwise. This position is close to that of *moral imperative,* or *duty.* But this debate over whether all decisions must come from the agents pursuing their own interest has to be left for further studies.

The second follows Simon's idea for the evaluation of choice models: choices are compared to those made by a human specialist. While we want to verify if choices are the same or not, this idea seems easy to implement. But if we want to argue that the artificial model chooses *better* than the reference human, we return to the problem of deciding what 'better' means.

The third candidate is some measure obtained from an aggregation of agents which are similar to the agent or behaviour we want to study. We so want to compare choices made by an agent based on some model, with choices made by some group to be studied (empirically, in principle). In this way we test realistic applications of the model, but assuming the principle that the decider agent represents in some way the group to be studied.

4.3 Combining the Two Approaches

A recent methodological approach can help us out here [20]. The phases of construction of theories are depicted in figure 1. However, we envisage several problems in the application of this methodology. Up front, the obvious difficulties in the translation from (T) to (E) and from (T) to (A), the subjectivity in the selection of the set of results (R) and corresponding observations (O), the formulation of hypotheses (H) from (E) (as Einstein said: "no path leads from the experience to the theory"). The site of the experimenter becomes again central, which only reinforces the need of defining common ground between him/her and the mental content of the agents in the simulation.

Thereafter, the picture (as its congeners in [20]) gives further emphasis to the traditional forms of experimentation. But Hales himself admits experimentation in artificial societies demands for new methods, different from traditional induction and deduction. Like Axelrod says: "Simulation is a third form of making science. (...) While induction can be used to discover patterns in data, and deduction can be used to find consequences of assumptions, the modelling of simulations can be used as an aid to intuition." [5, page 24]

This is the line of reasoning already defended in [13]: to observe theoretical models running in an experimentation test bed, it is 'exploratory simulation.' The difficulties in concretising the verification process (=) in figure 1 are even more stressed in [9]: the goal of these simulation models is not to make predictions, but to obtain more knowledge and insight.

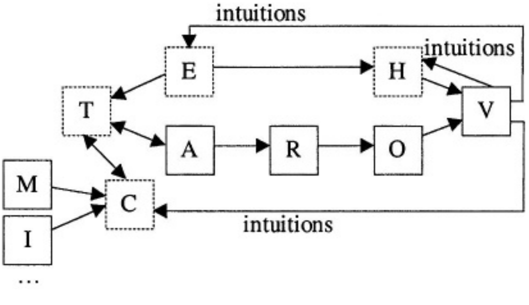


Fig. 2. Exploratory simulation. A theory (T) is being built from a set of conjectures (C), and in terms of the explanations (E) that it can generate, and hypotheses (H) it can produce. Conjectures (C) come out of the current state of the theory (T), and also out of metaphors (M) and intuitions (I) used by the designer. Results (V) of evaluating observations (O) of runs (R) of the program that represents assumptions (A) are used to generate new explanations (E), reformulate the conjectures (C) and hypotheses (H), thus allowing the reformulation of the theory (T).

This amounts to radically changing the drawing of figure 1. The theory is not necessarily the starting point, and the construction of explanations can be made autonomously, as well as the formulation of hypotheses. Both can even result from the application of the model, instead of being used for its evaluation. According to Casti [9], model validation is done qualitatively, recurring to intuitions of human specialists. These can seldom predict what occurs in simulations, but they are experts at explaining the occurrences.

Figure 2 is inspired in the scheme of explanation discovery of [20], and results from the synthesis of the scheme for construction of theories of figure 1, and a model of simulations validation. The whole picture should be read at the light of [9], that is, the role of the experimenter and his/her intuition is ineluctable. Issues of translation, retroversion and their validation are important, and involve the experimenter [10]. On the other hand, Hales' (=) is substituted by an evaluation machinery (V), that can be designed around values. Here, the link between agents and experimenter can be enhanced by BVG choice framework.

One of the key points of the difference between figures 1 and 2 is the fact that theories, explanations and hypotheses are being constructed, and not only given and tested. Simulation is precisely the search for theories and hypotheses. These come from conjectures, through metaphors, intuitions, etc. Even evaluation needs intuitions from the designer to lead to new hypotheses and explanations. This process allows the agent's choices to approximate the model that is provided as reference. Perhaps this model is not as accurate as it should be, but it can always be replaced by another, and the whole process of simulation can provide insights into what this other model should be.

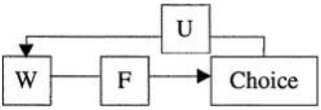


Fig. 3. Choice and update in the BVG architecture. W is the set of values the agent uses through choice function F to produce a decision. Then function U updates the system of values according to the assessment of the outcome of the previous decision.

In BVG (see figure 3), choice is based on the agent's values (say, machinery *W),* and performed by a function *F* (a set of control buttons, distributed by the *F* components). *F* returns a real value that momentarily serialises the alternatives at the time of decision. The agent's system of values is updated by a function *U* that uses multidimensional assessments of the results of previous decisions. We can represent the designer's choice model by taking these latter dimensions as a new set of values, W' (or, we can tune the function *F* manually, by moving those control buttons). Mechanisms *F* and *U* provide explicit means for drawing the link between the agent's (choosing) mind and the designer's experimental questions, thus transporting the designer into the (terms of the) experiment. This is accomplished by relating the backwards arrows in both figures (2 and 3). We superimpose the scheme of the agent on the scheme of the experiment.

5 Assessment of Experimental Results

This concern with experimental validation was an important keynote in the development of the BVG architecture. Initially we reproduced (using Swarm) the results of Axelrod's [4] "model of tributes," because of the simplicity of the underlying decision model. Through principled exploration of the decision issues, we uncovered certain features of the model previously unidentified by Axelrod. But the rather rigid character of the decision setting would not allow the model to show its full worth.

In our most elaborated experiment (in a Prolog test bed), agents select wines from a pool of options, in order to satisfy some (value-characterised) goals [3, 2]. This introduced new issues in the architecture, such as non-transitivity in choice, the adoption of goals and of values, non-linear adaptation, the confront between adaptation based on one or multiple evaluations of the consequences of decisions. We provide some hints into the most interesting results we have found.

In a series of runs, we included in F a component that subverts transitivity in the choice function: the same wine can rise different expectations (and decisions) in different agents. A new value was incorporated, to account for the effect of surprise that a slightly high price can raise, causing different evaluations (of attraction and of repulse).

The perils of subverting transitivity are serious. It amounts to withdrawing the golden rule of classical utility, that "all else being equal" we will prefer the cheaper option. However, we sustain that it is not necessarily irrational (sometimes) not to do so. We have all done that in some circumstances. The results of the simulations concerning this effect of surprise were very encouraging. Moreover, the agent's choices remained stable with this interference. The agent does not loose sense of what its preferences are, and what its rationality determines. It acts as if it allowed itself a break, in personal indulgence.

In other runs, we explored the role of values in regulating agent interactions, for instance, goal adoption. We found that when we increase the heterogeneity of the population in terms of values (of opposite sign, say), we note changes in the choices made, but neither radical, neither significant, and this is a surprising and interesting fact. The explanation is the "normalising" force of the multiple values and their diffusion. An agent with one or another different value still remains in the same world, sharing the same information, exchanging goals with the same agents. The social ends up imposing itself.

What is even more surprising is that this force is not so overwhelming that all agents would have exactly the same preferences. So many things are alike in the several agents, that only the richness of the model of decision, allied to their particular life stories, avoids that phenomenon.

The model of decision based on multiple values, with complex update rules, and rules for information exchange and goal adoption, presents a good support for decision making in a complex and dynamic world. It allows for a rich range of behaviours that escapes from directed and excessive optimisation (in terms of utilitarian rationality, it allows for "bad" decisions), but does not degenerate in pure randomness, or nonsense (irrationality). It also permits diversity of attitudes in the several agents, and adaptation of choices to a dynamic reality, and with (un)known information.

6 Conclusions

Whichever the experiment design, whoever conducts it, we don't think results can ever be considered absolutely valid. Never, in science. Just like a democratic election will never discover the best option, only the preferred one. The best choice does not exist, nor does the best criterion to decide it. What are the obtained results worth, then? Still, and despite all, the best possible, in each moment.

This is the essence of the calculus of importance (and democracy) behind the use of values. Even knowing importance is relevant, we have to accept that a calculus is possible and helps us, although in a field of uncertainty. So, every measure, every evaluation in complex phenomena is escorted by uncertainty rates. And still, we are here alive!

Accordingly, no prescribed methodology will ever be perfect for all situations. Our aim here is to draw attention to the role of the designer in any experiment, and also to the usually underaddressed issue of choice in the agent's architecture. Having a value-based choice model at our hands as a means to consider self-motivated autonomous agents, these two ideas add up to provide a complete decision framework, where the designer is brought into the experiment, through the use of common terms with the deciding agents.

References

1. Luis Antunes and Helder Coelho. Redesigning the agents' decision machinery. In *Affective Interactions,* volume 1814 of *LNAI.* Springer, 2000.
2. Luis Antunes, João Faria, and Helder Coelho. Choice: the key for autonomy. In *Proc. of EPIA 2001,* volume 2258 of *LNAI.* Springer, 2001. To appear.
3. Luis Antunes, João Faria, and Helder Coelho. Improving choice mechanisms within the BVG architecture. In *Intelligent Agents VII, Proc. of ATAL 2000,* volume 1986 of *LNAI.* Springer, 2001.
4. Robert Axelrod. A model of the emergence of new political actors. In *Artificial Societies – The Computer Simulation of Social Life.* UCL Press, 1995.
5. Robert Axelrod. Advancing the art of simulation in the social sciences. In *Simulating Social Phenomena,* volume 456 of *LNEMS.* Springer, 1997.
6. José Maria Castro Caldas. *Escolha e Instituições – uma análise económica apoiada na simulação multiagentes.* PhD thesis, ISCTE, Lisboa, 2000.
7. Cristiano Castelfranchi. Guarantees for autonomy in cognitive agent architecture. In *Intelligent Agents: agent theories, architectures, and languages, Proc. of ATAL'94,* volume 890 of *LNAI.* Springer, 1995.
8. Cristiano Castelfranchi. The theory of social functions: challenges for computational social science and multi-agent learning. *Journal of Cognitive Systems Research,* 2, 2001.
9. John L. Casti. Would-be business worlds. *Complexity,* 6(2), 2001.

10. Helder Coelho, Luis Antunes, and Luis Moniz. On agent design rationale. In *Proc. of the XI Brazilian Symp. on AI.* SBC and LIA, 1994.
11. Paul R. Cohen. A survey of the eighth national conf. on AI: Pulling together or pulling apart? *AI Magazine,* 12(1):16–41, 1991.
12. Paul R. Cohen. *Empirical Methods for AI.* The MIT Press, 1995.
13. Rosaria Conte and Nigel Gilbert. Introduction: computer simulation for social theory. In *Artificial Societies: the computer simulation of social life.* UCL Press, 1995.
14. Rosaria Conte, Rainer Hegselmann, and Pietro Terna. Introduction: Social simulation – a new disciplinary synthesis. In *Simulating Social Phenomena,* volume 456 of *LNEMS.* Springer, 1997.
15. António Damásio. *Descartes' error.* Putnam's sons, New York, 1994.
16. António Damásio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness.* Harcourt Brace, New York, 1999.
17. Nigel Gilbert. Emergence in social simulation. In *Artificial Societies: the computer simulation of social life.* UCL Press, 1995.
18. Nigel Gilbert. Models, processes and algorithms: Towards a simulation toolkit. In *Tools and Techniques for Social Science Simulation.* Physica-Verlag, 2000.
19. Nigel Gilbert and Jim Doran, editors. *Simulating Societies: the computer simulation of social phenomena.* UCL Press, London, 1994.
20. David Hales. *Tag Based Co-operation in Artificial Societies.* PhD thesis, Univ. Essex, 2001.
21. Steve Hanks, Martha E. Pollack, and Paul R. Cohen. Benchmarks, test beds, controlled experimentation, and the design of agent architectures. *AI Magazine,* 14(4), Winter 1993.
22. Stuart Russell and Peter Norvig. *Artificial Intelligence: A modern approach.* Prentice-Hall, 1995.
23. Stuart Russell and Eric Wefald. *Do the right thing – studies in limited rationality.* The MIT Press, 1991.
24. Herbert A. Simon. A behavioral model of rational choice. *Quarterly Journal of Economics,* 69:99–118, Feb. 1955.
25. Klaus G. Troitzsch. Social science simulation – origins, prospects, purposes. In *Simulating Social Phenomena,* volume 456 of *LNEMS.* Springer, 1997.
26. Philippe van Parijs. Functionalist marxism rehabilitated. A comment to Elster. *Theory and Society,* 11, 1982.
27. M. Mitchell Waldrop. *Complexity – The Emerging Science at the Edge of Order and Chaos.* Simon and Schuster, 1992.

Cognitive Identity and Social Reflexivity of the Industrial District Firms. Going Beyond the "Complexity Effect" with Agent-Based Simulations

Riccardo Boero[1], Marco Castellani[2], and Flaminio Squazzoni[2]

[1] Department of Economics, University of Pavia,
San Felice 5, 27100 Pavia, Italy,
rboero@eco.unipv.it

[2] Department of Social Sciences, University of Brescia,
San Faustino 74/b, 25122 Brescia, Italy,
{castella, squazzon}@eco.unibs.it

Abstract. Industrial districts (IDs) are complex inter-organizational systems based on an evolutionary network of interactions among heterogeneous, localized, functionally integrated and complementary firms. With an agent-based prototype, we explore how cognitive processes and social reflexivity dynamics of ID firms affect technological adaptation and economic performance of ID as a whole. Rather than observing IDs just by the point of view of the so-called bottom-up emerging properties, we try to study how firms develop over time "districtualized" behavioral attitudes, through cognitive capabilities of typifying and contextualizing in a social sense their technological, organizational and economic action. The question is: do cognitive processes, like those mentioned, have a great impact on technological learning and economic performance of firms over time?

1 Introduction

The paper aims to suggest an agent-based computational prototype able to investigate some micro-cognitive and social process underlying industrial districts (IDs). Rather than focusing just on what literature calls the "ID effect" (i.e., see: [16]; [42]), or looking mostly at what computational social scientists call "bottom-up emergent dynamics" ([20]), we have tried to focus on ID conceived as a cognitive state of mind of clustering complementary localized firms. The starting assumptions of the paper concern the idea that ID firms can behave in a different "districtualized" way ([5]). The principle is that more an ID firm is districtualized, more its behavioral attitude will be social oriented, while more it is dis-districtualized, more its behavioral attitude will be ego-centered. The districtualization of firms implies the strengthening of their tendency to consider the social context as an important source of information and of other relevant economic advantages that needs to be actively reproduced and to be taken into

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 48–69, 2004.

account as an important part of the individual decision. ID firms need to be conceived as cognitive agents that are able of developing over time "reflexivity capabilities" related to the capacity of typifying and internalizing the characteristics of their social context of experience as a stable structure, a positive part, as well as a reference of their cognitive individual action. This is what we mean for identification. More a firm develops over time a behavioral attitude actively committed to conceive the importance of the social context and to reproduce it through action, more it will identify itself with others and with ID as a whole.

In the traditional literature, the factors of complexity of IDs are related to the aggregate location dynamics emerging by the assumed homogeneous behavioral attitude of firms (i.e., automatic and natural commitment, cooperation, and trust among ID firms). The complexity is sought at the macro level of ID as a system, while ID firms behavior is transformed into a relatively simple black box. But, focusing on districtualization/dis-districtualization dynamics, cognitive individual processes of ID firms begin to be conceived as an important source of heterogeneity and complexity of IDs. Thus the complexity of IDs is also sought at micro level.

This is the reason why the so-called ID effect needs to be viewed not only as a mechanism that encapsulates the result of nonlinear dynamics emerging by interactions among heterogeneous, localized, complementary and functionally integrated firms, but also that includes a growth of complexity inside their cognitive identity processes. This is due to the fact that IDs, like any other social system, are not simply "swarm intelligence systems" with a composite, decentralized and un-intentional nature, but chiefly they are systems composed by social agents, that is to say by agents endowed with reflexive capabilities of monitoring, controlling, manipulating, typifying, making confidence toward environments which they move within, or in other words agents able to contextualize their action. They are social agents able to regulate actively their action according to factors such as the identification with others, the importance of the contexts, up to the confidence with the system as a whole. To understand IDs, two different levels of analysis need to be interlaced, namely that of interaction/nonlinear emerging dynamics/decentralized decision making of ID as a complex inter-organizational system, and that of reflexive typification/cognitive processes/social identity of firms as "social intelligent agents" embedded into the same industrial, spatial and social context

These are the standpoints of our theoretical approach on IDs, and this should be a point of convergence between recent debates on IDs, characterized by a micro-macro socioeconomic and cognitive perspective, and methodological debates on the use of agent-based models in social sciences. On one hand, our references come from the recent cognitive, network and evolutionary approach to IDs (i.e., see [1], [7], [26], and [40]), mostly on studies on "identity" and "identification" processes in IDs (see [41]). On the other hand, our references come from the field of multi-agent social systems and recent debates on "simple reactive and not cognitive" vs "reflexive and cognition provided" agents in agent-based social simulation (i.e., see [15]). Our opinion is that debates on "social intelligence" of artificial agents ([12], [13] and [32]), "micro-macro link" in agent-based models and needed levels of cognitive sophistication to simulate agents interacting within social settings ([14]), "first and second order emergence" ([24], [25]), and so on, have a great relevance for the possible integration of agent-based computational models as a part of the estate of social sciences. But our belief is that the challenge of modeling social phenomena can not be satisfied simply by a close application of "swarm intelligence" analogies (i.e., see [10]). Computational social scientists need to bring modeling to account with higher levels of complexity of social settings with respect to the famous case of ants and colonies. Our opinion is that such higher levels of complexity of social phenomena modeling are caused mostly by cognitive differences in reflexive capabilities developed by social agents in respect to non-human agents.

In conclusion, the paper shows an example of way of using agent-based simulation in the field of IDs to study, from a theoretical point of view, the role of social reflexivity capabilities of ID firms viewed as cognitive processes and carriers able to foster technological adaptation and economic performance of IDs. The second section shows how ID prototype works, from the point of view of its structural properties, that is to say classes of firms, division of labor among them, spatial localization of firms, and evolution of the technology and the market environment. The third section shows how ID cognitive agents work and by which computational building blocks they are composed. Blocks refer to what we call "information/action loop", which is a general theoretical framework able to reproduce computational cognitive processes undertaken by ID firms. The fourth section shows the analysis of simulation outcomes, with a focus on the most relevant emerging dynamics, above all, the relation among changing behavioral attitudes, technological learning, and economic performance of firms over time. Finally, the fifth section concludes with our intentions regarding the further development of ID prototype.

2 How ID Prototype Works

Speaking about an ID prototype[1] means to translate a general and abstract representation of an ID "archetype" into an agent-based computational architecture[2]. First of all, we start from a very broad and accepted definition of what

[1] The ID computational prototype has been created using Swarm libraries and Java programming language. Swarm is a toolkit for agent-based computational simulation developed by Santa Fe Institute (see: www.swarm.org) and used by a growing community of social scientists. For descriptions and applications of Swarm to economic phenomena, see [44], [31], and [30]. For all the details describing the ID prototype structure, see also [43]. To obtain the codes of the simulation, please write to one of the authors.

[2] The term "computational prototype" means that we have not modeled neither a specific and real ID, where the modeling operations aim to reproduce the reality in a more or less exhaustive way, nor an "abstraction" concerning fundamental mechanisms of social systems, where the modeling operations attend to study some

an ID archetype is: *an ID is a decentralized complex system characterized by an evolutionary network of interactions among heterogeneous, localized, functionally integrated and complementary firms.* Firms are embedded within a specific geographical area, they produce one-product goods for the market according to a division of labor based on production fragmentation, complementarity and mechanisms of coordination and integration of firms (i.e., see [4], [17] and [39]). Firms have rich proximity relations, both spatial and organizational oriented, and they move within specific technology and market environments.

Right from the start, we assume that ID agents are firms[3]. Firms are 400, divided in two different classes: final firms, having functions of organizing production and selling goods for the market, and sub contracted firms, having specialized functions related to the whole production process. The class of sub contracted firms is further divided into three sub-classes, sub firms A, B and C. It is well-known in the literature on IDs that final firms have a focal, strategic and innovative role. They have an interstitial position at the edge of market and ID, and they are the only ID agents having a vision of the production process as a whole (i.e., see [2], [9], [27] and [28]).

In order to produce goods for the market, firms interact giving rise to production chains, (i.e. partnership relations). We assume that a production chain must be composed of: 1 final firm + 1 sub firms A + 1 sub firm B + 1 sub firms C.

Firms have three basic features: technology (input), organizational asset (throughput), and economic performance (output). The relation among such three basic features is shown in figure 1, where the evolution of technology and market environment through which ID firms need to adapt is shown. Firms need to undergo 2000 simulation/production cycles, during which they face three phases of technological continuity and two phases of technological discontinuity. In short, over time, the market causes two technology breaking off (cycle 500 and 1000). In our sense, market is conceived as an "institution" collecting and distributing information about the performance of firms, technology evolution and consumers' needs.

Firms need to develop "absorptive capabilities" on technology and to learn the ways of adapting their organizational assets, trying to reach the fixed best technological practice level ([11]). We assume that technology (T1, T2, T3) implies an investment of internal organizational factors. Technology is composed

properties of social phenomena, as it happens, for example, with the competition-cooperation models, game theory models, and so on. The paper concerns an investigation of some theoretical mechanisms that are seeing in action within a common family of phenomena, namely that of IDs, through their synthesis in an ideal-type model.

[3] Clearly, the identification of agents ("computational units") and firms is a strong assumption, but it is a standard practice both in the literature on agent-based models and in the tradition of evolutionary economics applied to industrial business economics. Therefore, such reduction seems even less "strong" than ever in the case of IDs, since there is a close identification between management and entrepreneurship of firms (i.e., see [22] and other simulation ID models: [21]; [35]).

by a set of four numbers (e.g. 0, 3, 7, 2) (here, reference is [3]). Every number can be viewed as an organizational factor, such as labor, physical capital, human capital, and information and communication internal architecture. Firms start the simulation from a combination as follows: T1 (-1, -1, -1, -1), i.e. a situation of complete ignorance about technology factors. The best technological practice level of T1 is randomly fixed at the start of the simulation, and that of T2 and T3 is randomly fixed over time. This implies that firms can improve their technological effectiveness, both decreasing or increasing numbers/factors. Firms do not know the "best technological practice level" and can change numbers/factors just by turns. In short, firms deal with experimental day-to-day technological learning. Therefore, we assume that the experimental learning of firms is characterized by path dependence. The technological innovation of firms is affected by the technological position they have. In short, when firms take technological jumps from T1 to T2, or from T2 to T3, they start to explore the new combination of numbers/factors on the basis of their previous combination (i.e., previous combination: T1 3, 4, 7, 8/jump from T1 to T2/initial combination: T2 3, 4, 7, 8). According to the effectiveness of their organizational assets, in terms of distance/nearness of their combination of numbers/factors with respect to the best technological practice level, firms have specific costs and reach specific performance levels, as shown in table 3.

To adapt step-by-step their organizational asset, firms have two strategies of experimental exploration within the state of technological possibilities: "radical innovation" (with a possibility fixed on 80% to obtain a new value, namely a new number/factor); or "imitation by exploitation" of information coming from neighborhood (firms are able to look into the combination of factors of neighboring firms, to compare specific numbers/factors, to discover possible differences, and to imitate them) (here, our references come from evolutionary economics: [18], [33], and [37]).

Firms are located within an environment populated by other firms, namely the ID, with spatial neighborhood positions. The concept of neighborhood calls for the problem of proximity relations among firms. We introduce different metrics of proximity, viewed as different sources of information for firms. Over time, and with respect to "behavioral attitudes" of agents described afterward, firms develop a dynamic overlapping web of proximity relations with other firms, namely spatial, organizational and social forms of proximity (see: [6]; [45]). As it will be outlined in the next paragraph, proximity matters because it enables as byproducts sources of information, possibility of monitoring of the social context, and possibility of comparing individual characteristics and social context features. Proximity can be spatial enlarged, geography-dependent, organizational relation-dependent, or social-oriented.

To regulate all such computational operations, we introduce three tables, called *Change Matrix, Info Matrix,* and *Tech Matrix* (see table 1, 2 and 3,) where all actions are transformed in costs and values.

Finally, we introduce a double metrics of the firm profit. Firms have their individual level of profit, due to the difference between costs and levels of eco-

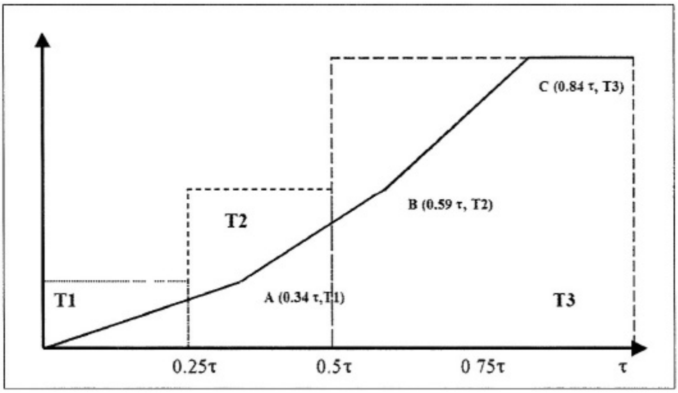


Fig. 1. Evolution of technology and market environment: τ is the number of production/simulation cycles. T1, T2 and T3 are the three technological regimes impacting ID firms over time. Phases of technology breaking-off are about cycles 500 and 1000. Bounded areas show technological positions and related achievable performance levels of firms with respect to technology standard and market evolution. We assume that technological evolution is irreversible (from T1 to T3) and characterized by growing level of cost and performance contents.

Table 1. *Change Matrix* shows costs needed to implement a new technology (first line) or to improve organizational asset, that is to say to change numbers/factors combination (second line). Along the column, there are all the three technological regimes impacting firms over time. Costs gradually increase over time with the gradual growth of market requests and performance needed.

	T1	T2	T3
Technological Change		200	400
Technical Change	50	100	200

Table 2. *Info Matrix* shows costs which firms must pay in order to achieve different type of information. Information concerns both technological strategies (innovation and imitation), and partnership selection mechanisms. The second case refers to different information criteria by which final firms organize their production chains, aggregating a team of sub contracted firms. Final firms continuously need information about economic, technology and organizational features of sub contracted firms in order to choose between stabilizing or destabilizing their inter-organizational contexts (chains).

	T1	T2	T3
Technology Imitation	40	70	
Organizational Asset Imitation	30	20	10
Technology Innovation	100	250	
Organizational Asset Innovation	80	50	30
Best Sub	5	5	5

Table 3. *Tech Matrix* shows data about costs and performance of firms in all the different learning steps undertaken by firms. As it is mentioned above, technology costs and economic performance gradually increase as well as the market requests over time. Column A shows technology costs, B shows levels of achievable performance, and C shows decreasing costs for the use of the same combination of numbers/factors for more than one simulation/production cycle. All costs and performance values are expressed by a continuum between worst and best technological practice levels, with an average calculus on the degree of distance/nearness of the combination of numbers/factors implemented by firms with respect to the best and worst levels.

	T1			T2			T3		
Organizational Asset	A	B	C	A	B	C	A	B	C
Worst	5	6	0.01	6.65	9.12	0.01	8.86	13.87	0.01
Best	7.32	10.49	0.01	9.74	15.96	0.01	12.97	24.26	0.01

nomic performance, as shown in column A and B of "Tech Matrix". But, at the aggregate level of chains, the total profit, generated by the product selling, is not the simple sum of the individual profits of interacting firms. We introduce an "extra profit" which mirrors the technological compatibility level of firms involved in the same chain. In order to produce quickly and to reach the highest possible level of quality of the goods on the market, firms need to "speak" the same technological language. In short, such extra profit emerging by the production-oriented aggregation of firms, is what we call, in our computational codes, the "time compression" value. The principle is that, because of fragmentation and the complementarity of ID production process, interacting firms need to be technologically compatible to produce high quality products at the right time.

3 How ID Agents Work

The foundation of the cognitive architecture of ID agents is based on the hypothesis that agents are able to process information about technology and the market environment, ID context, and organizational and economic features, and to translate it into a course of action, using what we call the "information/action" loop. It is a continuous loop which relates data to rough indexes, rough indexes to macro indexes, macro indexes to evaluations, and evaluations to actions. We set up an information set with different data concerning "day-to-day" activities of agents and different cognitive steps through which agents use, monitor and transform information into decisions. Such data is built on both temporal and spatial dimensions, and even on their interrelation.

We assume that computational capabilities of agents are bounded, and that time, memory and attention are finite and selective resources. We assume that agents can not act cognitively with parallel processing mechanisms, namely they can not control, manage and face the entire set of information with the same level of cognitive attention ([34]). Moreover, we assume that there is a trade-off between width and depth of the cognitive process. As underlined afterward, all cognitive steps undertaken by agents imply an information processing activity based on approximation, abstraction and synthesis of the relevant attributes belonging to information. In fact, the information/action loop starts with domain specific information and ends, using specific cognitive procedural processes, with broad generic information upon which the decision process is based.

The *first cognitive step* of the ID agent is the transformation of information into rough indexes of attribution. Information concerns the topics faced by firms. Rough indexes allow agents to assign a positive or negative judgment to information, which is expressed by a computational dichotomy of 0 and 1 values. Agents cluster, synthesize, and categorize information belonging to the same topics, transforming numbers into evaluations, through a first inference. Rough indexes are as follows:

1. "sold" (a first inference on market effectiveness of firms and their neighborhoods and a comparison among such values)
2. "time compression" (a first inference on technological compatibility of production chains and their neighboring chains, and a comparison among such values);
3. "performance" (an inference on the effectiveness on the market and a comparison with the neighborhood);
4. "number of chains" (an inference on the degree of stability and good relations among firms);
5. "selling firms" (an inference of the effectiveness of the system as a whole);
6. "technological change" (an inference of the degree of technological instability of the system as a whole);
7. "searching for new sub firms" (an inference of the instability of inter-firm relations and the tendency to the emergence of new partnership assets within the system as a whole);
8. "technology" (a comparison among the technology level of firms and neighborhood);
9. "organizational asset effectiveness" (a comparison between the level of effectiveness of organizational assets of firms and neighborhood);
10. "homogeneity of criteria for keeping sub firms" (an evaluation of the degree of uniformity of the inter-organizational assets within the neighborhood);
11. "homogeneity of criteria for searching sub firms" (an evaluation of the degree of diffusion of changes in the inter-organizational assets within the neighborhood);
12. "profit over time" (a comparison among levels of profit of firms and neighborhood over time, namely using an inference with temporal retrospective dimension);
13. "resources over time" (a comparison among levels of resources of firms and neighborhood over time, namely using an inference with temporal retrospective dimension);

14. "performance over time" (a comparison among performance values of firms and performance of neighborhood over time, namely using an inference with temporal retrospective dimension);
15. "investment on technology over time" (a comparison among average values of the technology investment of firms and neighborhood over time, namely using data of the last 20 simulation cycles);
16. "investment on organizational asset over time" (a comparison among average values of the investment on organizational assets of firms and neighborhood over time, namely using data of the last 20 simulation cycles).

The *second cognitive step* is the extrapolation of five macro aggregated indexes on previous rough indexes. Agents develop a level of cognitive abstraction that is more synthetic compared to the previous step. Macro indexes are:

a) *partnership* (a macro inference of the positive or negative nature of the features of the partnership context),
b) *environment* (a macro inference of the stable or unstable nature of the technology and the market environment),
c) *technology* (a macro inference of the individual degree of technological effectiveness),
d) *organization* (a macro inference of the positive or negative nature of the organizational fundamentals of the firm),
e) *economic* (a macro inference of the positive or negative nature of the economic fundamentals of the firm).

Macro indexes synthesize and cluster rough indexes at a higher level of cognitive abstraction. The relation between macro indexes and rough indexes conforms to the following general rule: $M_a = W_{a1}R_{a1} + W_{a2}R_{a2} + ... + W_{an}R_{an}$ where M_a , W_{a1} , R_{a1} , etc.. $\epsilon[0, 1]$ and M_a represents a macro index, W_{a1} is the weight of the first rough index R_{a1} and so on. Thus, the shift from rough to macro indexes is based on a simple computational procedure, a weighted average, performed by agents over time. Such a procedure is characterized by a heterogeneous assignment of relevance-driven attention undertaken by agents on specific indexes.

The *third cognitive step* is the indexes evaluation upon which differences in the behavioral states of agents may or may not matter. According to the information/action loop, before acting, agents need to be able to evaluate such indexes. The evaluation process calls for the problem of which kind of "behavioral attitudes" agents develop over time. Attitudes are conceived as different possible states of agent behavior emerging from a continuum between degrees of districtualization. Agents can be more or less districtualized, in the sense that their behavior can be affected by the characteristics of their social context. An agent less districtualized is pushed to think more in terms of "individual centered self". It can set aside the features of its context of interaction and social experience. Its decisions are not particularly bounded by social neighborhood influences. Otherwise, an agent more districtualized is pushed to think more in terms of "social group self" (i.e., see [36]). Its attitude is characterized by a more active identification with other agents belonging to the same context of experience. In short, we assume that features of the social context have a deep influence on the individual cognitive process when social reflexivity of agents grows over time. For social reflexivity, we mean the capacity of an agent of typifying and internalizing the characteristics of its social context of experience as a stable structure, a positive part, as well as a reference of its cognitive individual action.

The possible behavioral attitudes of an agent are *state 0,* or what we call "self-centered" attitude (agent is located in a context, it has specific neighboring agents; it enacts production relations with other agents, produces and sells products, trying to increase its economic performance, its technological profile, its organizational asset, and so on; it is not interested in establishing stable and rich relations with other agents, that is to say it seeks one-shot interactions, focusing continuously on imperatives of the economic performance), *state 1,* or what we call "chain-management" attitude (agent is interested in maintaining stable and rich relations with other interacting agents; it thinks about the chain as an unit, i.e. a locus of organizational relations and relevant information and a source of technological learning coordination; it starts to conceive complementary relations with others, and enlarges its "state 0"-context to other agents, with both spatial and organizational neighborhood), *state 2,* or what we call "clustering" attitude (agent with the chain management attitude meets other agents with the same attitude which put trust on the importance of what can be called policies for the social horizon enlargement; agent belonging to stable chains enlarges its microcosm to other agent belonging to stable neighboring chains; it exchanges information with other agents without having direct interactions), *state 3,* or what we call "grouping" attitude (agent starts to reflect upon the collective properties of the cluster trying to improve the collective effectiveness of the cluster; it recognizes all the other agents as members of the cluster and interacts with them, eventually exchanging information and partners within the group and making social distribution policy of the extra profit).

Firms start the simulation as self-centered attitude agents (state 0). We assume that all shifts among states depend on the following rule: whenever agents develop the perception of possible economic benefits which can emerge by cooperation with others, then they are pushed to define better, and in a more stable way, their contexts of interaction. Agents start to conceive their context of interaction as a tool of learning and information. Exploring the context, agents create with others a kind of relational tie, where information is exchanged, learning takes mutual directions, and resources can be shared. As it will be outlined, this implies that agents develop a cognitive representation of their tasks in terms of "relationship" ([8]).

We define the agent-based process of elaboration and change of behavioral states from the bottom-up (from state 0 to state 3) over time as *morphogenetic shifts.* The shift from state 0 to state 1 depends on the emergence of a relative stability of production chains (five cycles of recurring interactions with the same team of sub contracted firms) and on specific conditions of the macro indexes on "partnership" and "economic fundamentals" (macro index of partnership >= 0.75; macro index of economy >= 0.75). According to such conditions, agents can develop a behavioral attitude toward the transformation of previous recurring interactions into stable partnership relation, loosing their previous self-centered attitude. We assume that agents, facing a state of good economic performance and perceiving a potential good context of interactions, are pushed to define, in a more binding way, their organizational relations. In a sense, agents put trust in their organizational neighborhood contexts. The shift from state 1 to state 2 depends on a specific condition of the "partnership index" (value of 0.95). A clustering behavioral attitude implies the interest about, and a sharing of the information contained within, the whole spatial neighboring firms. The next step is the diffusion of the clustering behavioral attitude within spatial neighboring chains, when a similar condition of trust in the partnership mutually grows among agents. This is the mechanism which allows the diffusion of state 2 among firms. This implies that spatial proximity relations start to develop cluster proximity relations. The shift from state 2 to state 3 depends on conditions as follows: if "partnership index" or "environment index", or both are > 0.75, then at least two of three others ("technology", "organization" and "economic index") >= 0.75; if the spatial neighboring agents are already in state 3. "Partnership index" and "environment index" give agents the trust on the positive global state of the industry as a whole. The other indexes show a positive combination related to the individual state of the agent. We assume that, in this condition, agents are pushed to reflect in a more global way and to conceive the problem of the relation between individual effectiveness and the collective effectiveness of the group as a whole.

Bottom-up shifts are not equated with linear and irreversible processes. In fact, agents can develop, change and destroy continuously their behavioral attitudes, over time, by means of mechanisms of behavioral attitude *deconstruction shifts.* It is a matter of cognitive adaptation in respect to contexts and environments. Facing some "positive" cognitive configuration, agents develop a bottom-up process of elaboration of their behavioral attitudes (from 0 to 1, and so on), while facing some "negative" cognitive configuration, agents destroy their behavioral attitudes turning back to previous steps, in a top-down process.

The process of deconstruction of the behavioral attitudes conforms to a general computational rule as follows: if all the macro-indexes, both external and internal, are <= 0.5, then agents shift from state 1, 2, or 3 to state 0. Moreover, we assume other specific conditions for deconstructing behavioral attitudes. The deconstruction from state 3 to state 1, or the *group exit option,* if "technology", "organization" and "economic index" < 0.25, and if "partnership" and "environment index" > 0.5; deconstruction from state 2 to state 0, or the *cluster exit option,* if production chain asset is broken (a production chain is broken if one of these four conditions is true: the product has not been sold, $\text{profit}_t < \text{profit}_{t-5}$, $\text{resources}_t < \text{resources}_{t-5}$, $\text{time compression}_t < \text{time compression}_{t-5}$); and deconstruction from state 1 to state 0, or the *free hands option,* if production chain asset is broken (a production chain is broken for the same reasons as above).

Therefore, agents develop different behavioral attitudes over time, and they act in different operation fields having finite action recipes, as shown in table 4 and 5. The operation fields are what we call *"technology", "keep", "search",* and *"share"*. "Technology" refers to the need of agents to exploit context-based local information to improve their technology and organizational assets. "Keep" and "search" refer to how agents manage their partnership relations, at the edge of needs of stabilization and thrusts of de-stabilization of their relational contexts. "Share" refers to the policy of chain profit management that agents conduct. Behavioral attitudes have properties to relate such fields to specific action recipes. As it is shown in table 4 and 5, different behavioral attitudes imply the use of specific action recipes in specific operation fields.

At the start of the simulation, agents use a specific recipe assigned in a random way. Over time, they carry on using it, by transforming the recipe into routine. The routine is broken when macro indexes push the agent to change it. Therefore, the agent starts a phase of trial and error processes trying to define a new routine within action recipes. Thus, routines can be maintained or changed, and this is a focal phase of the agent action.

The role of macro indexes and their configuration is fundamental for understanding why and how agents change or maintain their routines. Macro indexes configuration is conceived as the adaptation mechanism that forces agents toward learning about routines. We set a fixed number of indexes configurations and the presence of a kind of *ringing bell* mechanism which represent the capacity of agents to perceive the presence of an unsatisfactory routine. Specific configurations of macro-indexes cause the activation of the ringing bell mechanism, driving the attention of the agent on a specific topic. The ringing bell mechanism means that the agent has some problem with its routines. It is based on the hypothesis that the selective attention of agents is oriented toward fixed operation fields, and directed to specific significant areas of the problem space, by means of a sort of *distinctiveness* ([29]).

According to such fixed combinations, and because of cognitive limitations about memory, time, attention and self-monitoring, the agent can change its routines within a specific operation field (technology, keep, search, share). It has limitations in proceeding to estimate the routines value (goodness) and in identifying the "critical" routine. The agent starts to perceive a problem on a specific operation field and develop a phase of evaluation and learning based on the exploration of other possible action recipes based on a memory function that collects data on the last five times period where a specific routine has been used. The agent needs to learn how to solve the problem within an operational field defining new routines. The ringing bell mechanism works, in the case of sub firms, if the "technology index" and the "economic index" < 0.25, the ringing bell focuses on "technology"; the agent perceives the necessity of change its routine on such operational field, while, in the case of final firms, if all the indexes $<= 0.5$, the agent falls into a "panic condition" and starts to change randomly its routines in one or two different operation fields, otherwise, if the "technology index" < 0.25, the agent starts to change its routine in the "technology" operation field; if the "organization index" < 0.25, the agent starts to change its routine with equal probability in the "keep" or in the "search" operation fields, while if the "organization index" is < 0.25 or >= 0.5, the agent starts to change its routine in the "share" operation field; and if the "economic index" < 0.25, the

Table 4. Relations among "operation fields", "behavioral attitudes" and "action recipes". Part 1.

Action Recipes	**Behavior Attitudes**	**Operation Fields**
look at the first agent with different technology/organizational asset you meet	Self Centered	Technology imitation in the sub-fields of technology and organization asset
look at the first agent with different technology/organizational asset you meet, which has sold its product		
look at the agent with different technology/organizational asset you meet, which has a percentage of extra-profit better than yours and the highest available		
look at the agent with different technology/organizational asset you meet, which has a behavioral attitude higher than yours and the highest available	Chain Management Clustering	
look at the agent with different technology/organizational asset you meet, which has a level of resources better than yours and the highest available		
look at the agent with different technology/organizational asset you meet, which has a level of profit better than yours and the highest available		
look at the agent with different technology/organizational asset you meet, which has a level of cost higher than yours and the highest available	Grouping	
look at the agent with different technology/organizational asset you meet, which has a level of effectiveness of organizational asset better than yours and the highest available		
look at the agent with different technology/organizational asset you meet, which has a level of investment on technology/organizational asset better than yours and the highest available		
look at the first agent with different technology/organizational asset you meet, which has a level of performance better than yours and the highest available		

agent starts to change its routine in the “technology” operation field and with equal probability its routine in the “keep” or in the “share” operation field. The

Table 5. Relations among “operation fields”, “behavioral attitudes” and “action recipes”. Part 2.

keep your team of sub firms if time compression $\Delta_{t,t-1} >= 0$	Self Centered	Keep strategy of partnership stabilization
keep your team of sub firms if profit $\Delta_{t,t-1} >= 0$		
keep your team of sub firms if resources $\Delta_{t,t-1} >= 0$		
keep your team of sub firms if you have sold your product	Chain Management Clustering Grouping	
keep your team of sub firms if time compression $\Delta_{t,t-5} >= 0$		
keep your team of sub firms if profit $\Delta_{t,t-5} >= 0$		
keep your team of sub firms if resources $\Delta_{t,t-5} >= 0$		
search for a new team of sub firms randomly	Self Centered Chain Management Clustering Grouping	Search strategy of partnership definition
search for a new team of sub firms focusing on who has the highest investment on organizational asset		
search for a new team of sub firms focusing on who has the highest performance		
search for a new team of sub firms focusing on who has the most similar technology and organizational asset configuration		
give to your partners the 0% extra-profit	Self Centered	Sharepolicy of chain extra-profit management and distribution
give the 5% extra-profit to each partner		
give the 10% extra-profit to each partner		
give the 13.3% extra-profit to each partner	Chain Management Clustering	
give the 16.6% extra-profit to each partner		
give the 20% extra-profit to each partner		
give the 23.3% extra-profit to each partner	Grouping	
give 25% extra-profit to each partner		
give the 70% extra-profit to partners, distributed proportionally according to their needs		
distribute proportionally the 100% extra-profit according to the needs of each member of the chain		

principle of changing routines is that an agent facing the perception of a problem within an operation field can use its memory on past routines, that is to say the last five periods of time during which a routine has been used, to support its routine definition process. Agent can relate routines to macro-indexes in order to define positive or negative associated values. According to the memory function, it changes, evaluates and chooses routines[4].

In conclusion, the cognitive architecture of ID agents is based on cognitive typification activities which relate continuously individual experience and social contexts. Macro-indexes evaluation is a cognitive step through which agents try to incorporate information and to develop "attribution" about the state of technology and the market environment, characteristics of their relational contexts and their own individual features, in order to find appropriate strategies of technological learning. Reflexive typification works through the capacity of agents to assign objective characteristics both on their experiences, their social context, and their operation environment. Agents are able to elaborate day-to day ordered evaluations about what they have done, and day-to-day monitoring evaluations about which kind of social context they are moving in. In a sense, action has here what Emirbayer and Mische call a "practical-evaluative dimension" associated with a "relational dimension" ([19]).

4 Analysis of Outcome and Emerging Dynamics

If we observe the simulation outcome[5], we can sketch several inferences. Technological breaking-off phases imply a selection of ID firms, even with different

[4] Here, the cognitive process of routine definition is based on the following steps: the ID agent has memory of routines implemented in the past, even if concentrated upon macro-indexes and bounded to some time periods (last five cycles of implementation); its space of possible routines is limited by its behavioral attitude, as it is shown in table 4 and 5; the agent uses continuously memory function for developing data about all routines used; if within the space of all the possible routines, there is a routine not yet explored, the agent chooses that one; the agent creates an average value of data collected on past routines; in the case of complete exploration of all possible routines, using data referring to the past, the agent defines its new routine according to an evaluation about the relation between routines and macro-indexes.

[5] To test ID prototype, we use several indicators. By observing them, it is possible to grasp fundamental dynamics emerging by ID prototype. We have also created different simulation settings in order to reinforce evidence about how behavioral attitudes and the typology of social contexts affect the performance of firms (running the ID prototype it is possible to choose all the different combinations of behavioral states, right from the start of the simulation). The indicators we use here are as follows (running the model, it is possible to observe and produce other kinds of indicators): a) final firms matching market requests over time; b) final firms performance and behavioral attitudes over time; c) final firms performance in different prototype settings running separately with behavioral attitudes state 0, state 0 and 1, and with complete behavioral states; d) weight of the different macro indexes over time; e) dimension of the neighborhood relations over time; f) technological adaptation level over time.

dynamics. As it is shown in figure 2, the first discontinuity phase (about cycle 500) is absorbed by 88% of the firms, while the second phase causes a strong oscillation in the firms performance, but without implying a further exit of firms from the market. This is due to the fact that firms over time are more effective in technological learning, despite the growth of costs and request of the technological quality of their goods marked by the market.

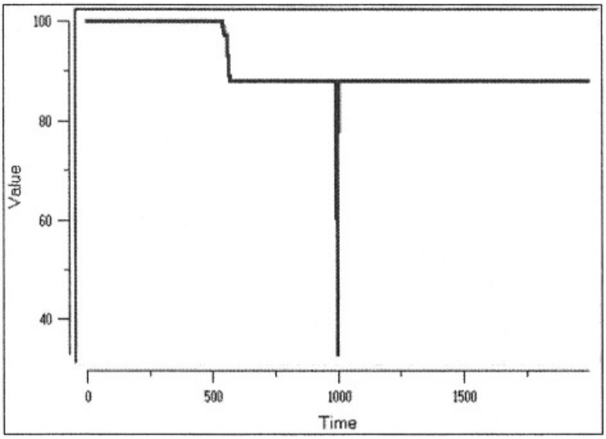


Fig. 2. Final firms matching market requests over time.

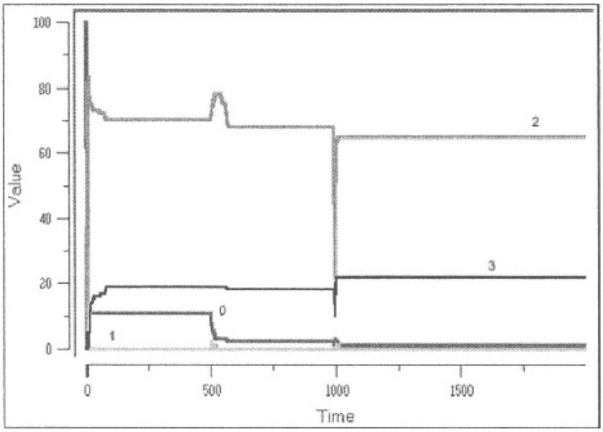


Fig. 3. Final firms matching market requests over time at variance of behavioral attitudes.

The evolution of behavioral attitudes states over time shows that firms facing technological discontinuity and increasing market pressure phases develop different strategies of response over time, while firms facing technological continuity tend to stabilize their behavioral attitudes. As it is shown in figure 3, the first phase of technology and market stability (until 500 cycle) shows a tendency of agents to lock-in their behavioral attitudes with many in state 2 ("clustering attitude") and few of them in state 0 ("self centered attitude"). The 10% of agents are quickly able to develop the "grouping attitude" (state 3), while another 10% of agents lock-in their behavioral attitude right from the start in the state 0. Such stabilization of the behavioral attitudes goes on until the first technology breaking-off (around cycle 500). In this phase, agents in more critical technology and market conditions try to develop their behavioral attitudes, above all shifting from state 0 to state 2, but without success. They are the first and the only victims of the market selection. The behavioral state 1 ("chain-management attitude") is just a shelter in times of technology and market deeper challenge, along all the simulation time. Just as before, the second phase of technological stability shows a long *durée* settlement of behavioral attitudes of agents.

Certainly, the second phase of technological instability is more interesting than the previous one (around cycle 1000). As it is shown in figure 2, here ID firms go through a deep but quick phase of market crisis. How do firms face such crisis, from the point of view of their behavioral attitudes? As it is possible to observe comparing figure 2 and 3 around cycle 1000, firms involved in such crisis are above all those in state 2 ("clustering attitude"). They do not simply destroy their behavioral state, for instance passing from state 2 to state 0, but some firm develops a state 3-like behavioral attitude. It is worth noting that the so called "grouping" behavioral attitude of firms stands up despite the two technological and market crisis, and even strengthens over time.

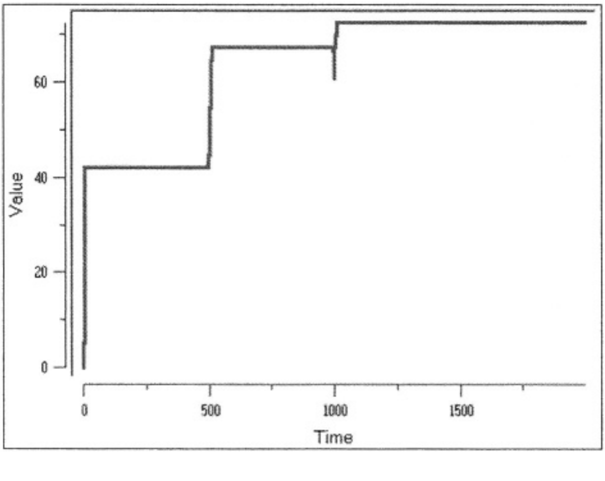


Fig. 4. Average dimension of neighborhood proximity relation.

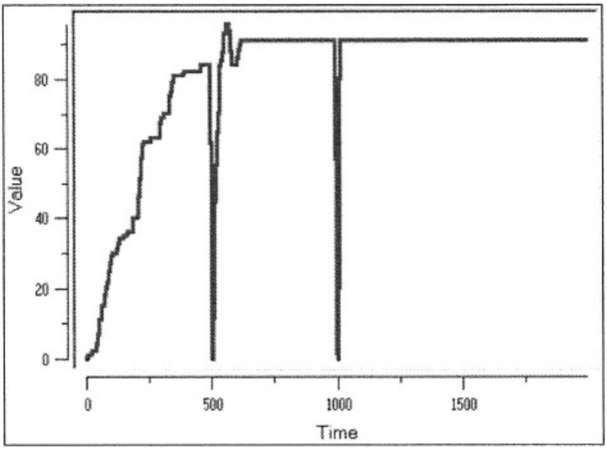


Fig. 5. Level of technological learning of final firms over time.

Such strengthening-effect of identification attitudes over time can be confirmed if one observes figure 4, where data about the dynamics of the average dimension of neighborhood proximity relations allows us to observe how large the context of relations enacted by agents over time is. Such enlargement of the social horizon of agents not only grows over time, but rather it grows during phases of technology and market adaptation challenges. Such data tells us that identification dynamics developed by agents are not only a fundamental tool of technological learning and economic performance for firms, but rather ID firms develop a polarization of grouping attitudes over time, and they are over time more enhanced and reinforced, when agents face technology and market adaptation needs. As it is shown in figure 5, the growing tendency of firms to stabilize the behavioral attitudes upward affects their technological learning level, and above all their capacity to experimentally discover the best technological level to stay on market.

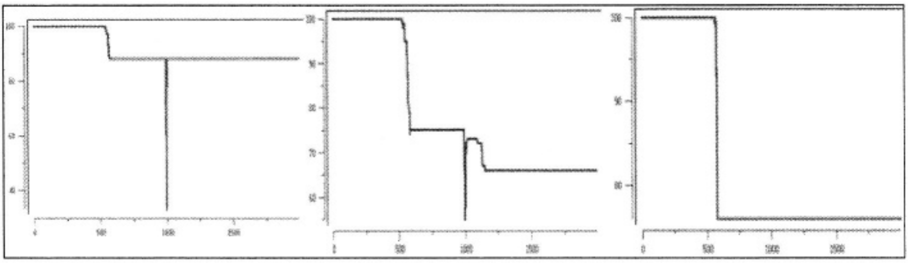


Fig. 6. Final firms matching market requests on different experimental settings. From the left to the right, outcomes of state "complete", state 0 and 1, and state 0.

Finally, we have created different prototype settings by changing mechanisms of behavioral attitudes development. As it is shown in figure 6, we set a prototype running just with behavioral attitude state 0, and running with state 0 and 1. The outcome of a set with state 0 and 1 confirms that behavioral attitude state 1 ("chain management attitude") causes the loss of the advantage of market-oriented "self centered attitudes" without generating the advantage of information source and processing typical of a wide relational context, as in the state 2 and 3. Figure 6 shows that at the end of simulation cycles, levels of firms still on market are as follows: 88% in the complete set, 76% in state 0, and 66% in state 0 and 1.

In conclusion, the simulation outcome of ID prototype shows that ID agents are able to develop different behavioral attitudes over time, such attitude development has a positive effect on the long-time period learning of agents, and social relational context and districtualized behavioral attitude are more deeply developed during phases of technology and market challenge. As in the case of technology breaking-off phases, if we compare figure 4, 5 and 6, it is possible to stress that districtualization of firms within their contexts of action are not a negative constraint upon the individual economic imperative, but rather a positive source of information and learning about environment challenges.

5 Conclusion: How to Further Develop ID Prototype

Our intentions are to test ID prototype by means of an empirical investigation that will be undertaken in some representative IDs in Italy. Such an investigation will be conducted through a qualitative questionnaire on ID firms. It will be devoted to the function of extracting relevant information on behavioral attitudes of ID firms. The investigation will concern all the mechanisms theoretically investigated by the ID prototype. It will be a test that will attend to simplify the assumptions on the model building blocks that have been described in this paper. Across the different typology of Italian IDs that has been already classified by several statistical surveys, which will give us a quite complete set of data on the structure and the performance of Italian IDs over time ([38]), according to a organizational spectrum that goes from relative traditional IDs to firm-centered network-like IDs, we will be able to choose some representative examples for all the different typology to investigate. The focus will be on cognitive processes of contextualization and identification and how they come to affect the decision of ID firms over time and on the relation between them and technological learning and market performance of firms.

Our belief is that this investigation should be a good way of opening the black box of ID firms and to focus on the question of complexity ex-ante and not only ex-post in the agent-based social simulation models. The second step will be to integrate the outcome of empirical investigations into our computational prototype, in order to deeply figure out the way to go for framing a complexity theory of IDs able to incorporate all the relevant achievements of social science theories in the field.

Acknowledgments. A preliminary version of the paper has been discussed in a workshop at the Department of Social Sciences (DSS), University of Brescia, Italy, 15th May 2002. We would like to thank people of DSS and Sergio Albertini for their useful remarks. Our paper, but really all our research program, owes much to ideas, illuminating insight, and support given us by Giancarlo Provasi (University of Brescia). Moreover, we would like to thank Pietro Terna (University of Turin), the core of the Swarm Community in Italy. Finally, we would like to thank the anonymous referees of RASTA 2002 and all the people that have attended the workshop, too. Their precious suggestions have been really useful to improve the quality of our paper. Obviously, we alone are responsible for the contents of the paper.

References

1. Albertini S.: Networking and Division of Labour: The Case of Industrial Districts in the North-East Italy, Human Systems Management, **18** (1999) 107–115
2. Albino V., Garavelli A. C., Schiuma G.: Knowledge Transfer and Inter-Firm Relationships in Industrial Districts: The Role of the Leader Firm, Technovation, **19** (1999) 53–63
3. Ballot G. and Taymaz E.: Technological Change, Learning and Macro- Economic Coordination: An Evolutionary Model, Journal of Artificial Societies and Social Simulation, vol. 2, **2** (1999) <http://www.soc.surrey.ac.Uk/JASSS/2/2/3.html>
4. Becattini G.: The Marshallian Industrial District as a Socio-economic Notion, in Pyke F., Becattini G., and Sengenberger W. (Eds.), Industrial Districts and Inter-Firm Cooperation in Italy, International Institute of Labour Studies, Geneva, (1990) 37–51
5. Becattini G.: From the Industrial District to the Districtualization of Production Activity: Some Considerations, in Belussi F., Gottardi G. and Rullani E. (Eds.), The Net Evolution of Local Systems. Knowledge Creation, Collective Learning and Variety of Institutional Arrangements, Kluwer, Dordrecht, forthcoming
6. Bellet M., Kirat T. and Largeron C. (Eds.): Approaches Multiforms de la Proximité, Hermes Science Publications, Paris (1998)
7. Belussi F. and Gottardi G. (Eds.): Evolutionary Patterns of Local Industrial Systems: Towards a Cognitive Approach to the Industrial District, Ashgate, Aldershot Brookfield Singapore Sydney, (2000)
8. Bickhard M. H.: Information and Representation in Autonomous Agents, Journal of Cognitive System Research, (2000) **1**
9. Boari C. and Lipparini A.: Networks within Industrial Districts: Organizing Knowledge Creation and Transfer by means of Moderate Hierarchy, Journal of Management and Governance, (1999) **3** 339–360
10. Bonabeau E., Dorigo M., Theraulaz G.: Swarm Intelligence. From Natural to Artificial Systems, SFI Studies in the Sciences of Complexity, Oxford University Press, New York, (1999)
11. Cohen W., Levinthal D.: Absorptive Capacity: A New Perspective on Learning and Innovation, Administrative Science Quarterly, (1990) **35** 128–152
12. Conte R.: Social Intelligence among Autonomous Agents, Journal of Computational and Mathematical Organization Theory, 5, (1999) **3** 203–228

13. Conte R.: The Necessity of Intelligent Agents in Social Simulation, in Ballot G., Weisbuch G. (Eds.), Applications of Simulation to Social Sciences, Hermes Science Publishing Ltd, Oxford, (2000) 19–38
14. Conte R. and Castelfranchi C.: Simulating Multi-Agent Interdependencics: A Two-Way Approach to the Micro-Macro Link, in (Eds.), Troitzsch K. G, Mueller U., Gilbert N. e Doran J., Social Science Microsimulation, Springer-Verlag, Berlin, (1996) 394–415
15. Conte R., Edmonds B. and Moss S., Sawyer R. K.: Sociology and Social Theory in Agent-Based Social Simulation: A Symposium, Journal of Computational and Mathematical Organization Theory, (2001) **7** 183–205
16. Costa-Campi M. T. and Viladecans-Marsal E.: The District Effect and the Competitiveness of Manufacturing Companies in Local Productive Systems, Urban Studies, Vol. 36, (1999) **12** 2085–2098
17. De Propis L.: Systemic Flexibility, Production Fragmentation and Cluster Governance, European Planning Studies, Vol. 9, (2001) **6** 739–753
18. Dosi G.: Innovation, Organization and Economic Dynamics. Selected Essays, Edward Elgar, Cheltelham, UK, (2000)
19. Emirbayer M. and Mische A.: What is Agency?, American Journal of Sociology, 103, (1998) **4** 962–1023
20. Epstein J. M., Axtell R.: Growing Artificial Societies. Social Science from the Bottom-Up, MIT Press, Cambridge, Massachusetts, (1996)
21. Fioretti G.: Information Structure and Behaviour of a Textile Industrial District, Journal of Artificial Societies and Social Simulation, vol. 4, (2001) **4** <http://www.soc.surrey.ac.uk/JASSS/4/4/1.html>
22. Fullet T. and Moran P.: Small Enterprises as Complex Adaptive Systems: A Methodological Question?, Entrepreneurship and Regional Development, (2001) **15** 47–63
23. Giddens A.: The Constitution of Society. Outline of a Theory of Structuration, University of California Press, Berkeley and Los Angeles, (1986)
24. Gilbert N.: Holism, Individualism and Emergent Properties. An Approach from the Perspective of Simulation, in Hegselmann R., Mueller U. and Troitzsch K. G. (Eds.), Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View, Kluwer Academic Publishers, Dordrecht / Boston / London (1996) 1–27
25. Gilbert N. and Terna P.: How to Build and Use Agent-Based Models in Social Sciences, Mind & Society, I, (2001) **1** 57–72
26. Lane D.: Complexity and Local Interactions: Towards a Theory of Industrial Districts, Curzio Quadrio A. amd Fortis M. (Eds.), Complexity and Industrial Districts, Springer Verlag, Berlin (2002)
27. Lazerson M. H. and Lorenzoni G.: The Firms that Feed Industrial Districts: A Return to the Italian Source, Industrial and Corporate Change, vol. 8, (1999) **2**
28. Lazerson M. H. and Lorenzoni G.: Resisting Organizational Inertia: The Evolution of Industrial Districts, Journal of Management and Governance, (1999) **3** 361 377
29. Lockhart R. S. and Craik F. I. M.: Levels of Processing: A Retrospective Commentary on a Framework for Memory Research, Canadian Journal of Psychology, (1990) 44 87–112
30. Luna F., Stefansson B. (Eds.): Economic Simulation in Swarm: Agent-Based Modelling and Object Oriented Programming, Kluwer Academic Publishers, Boston/Dordrecht/London, (2000)
31. Luna F., Perrone A. (Eds.): Agent-Based Methods in Economic and Finance: Simulations in Swarm, Kluwer Academic Publishers, Boston/Dordrecht/London, (2001)

32. Malsch T.: Naming the Unnamable: Socionics or the Sociological Turn of/to Distributed Artificial Intelligence, Autonomous Agents and Multi-Agent Systems, (2001) **4** 155–186
33. March J. G.: Exploration and Exploitation in Organizational Learning, Organization Science, 2, (1991) **1** 71–87
34. March J. G.: A Primer on Decision Making, The Free Press, New York, (1994)
35. Minerva T., Poli I. and Brusco S.: A Cellular Automaton as a Model to Study the Dynamics of an Industrial District, Paper presented to the workshop "Complexity and Industrial Districts", Modena and Reggio-Emilia University, Department of Cognitive, Social and Quantitative Sciences, December 2001.
36. Moran P.: Personality Characteristics and Growth-Orientation of the Small Business Owner-Manager, International Small Business Journal, 16, (1998) **3** 17–39
37. Nooteboom B.: Learning, Innovation and Industrial Organization, Cambridge Journal of Economics, (1999) **23** 127–150
38. Paniccia I.: One, Hundred, Thousand of Industrial Districts: Organisational Variety in Local Network of Small and Medium Enterprises, Organisation Studies, 19, (1998) **4**
39. Pyke F. and Sengerberger W. (Eds.): Industrial Districts and Local Economic Regeneration, International Institute for Labour Studies, Geneva, (1992)
40. Rullani E.: The Industrial District (ID) as a Cognitive System, in Curzio Quadrio A. and Fortis M. (Eds.), Complexity and Industrial Districts, Springer Verlag, Berlin (2002)
41. Sammarra A. and Biggiero L.: Identity and Identification in Industrial Districts, Journal of Management and Governance, (2001) **5** 61–82
42. Signorini L.: The Price of Prato, or Measuring the Industrial District Effect, Papers in Regional Sciences, (1994) **73** 369–392
43. Squazzoni F. and Boero R.: Economic Performance, Inter-Firm Relations and Local Institutional Engineering in a Computational Prototype of Industrial Districts, Journal of Artificial Societies and Social Simulation, vol.5, (2002) **1**, <http://jasss.soc.surrey.ac.uk/5/1/1.html>
44. Terna P.: Simulation Tools for Social Scientists: Building Agent-Based Models with Swarm, Journal of Artificial Societies and Social Simulation, vol.1, (1998) **2**, <http://www.soc.surrey.ac.uk/JASSS/1/2/4.html>
45. Torre A. and Gilly J.- P.: On the Analytical Dimensions of Proximity Dynamics, Regional Studies, Vol.34, (2000) **2** 169–180

The MAS-SOC Approach to Multi-agent Based Simulation

Rafael H. Bordini[1]*, Fabio Y. Okuyama[1], Denise de Oliveira[1],
Guilherme Drehmer[1], and Romulo C. Krafta[2]

[1] Informatics Institute
Universidade Federal do Rio Grande do Sul (UFRGS)
CP 15064, 91501-970, Porto Alegre–RS, Brazil
{bordini,okuyama,edenise,gdrehmer}@inf.ufrgs.br
[2] Faculty of Architecture
Universidade Federal do Rio Grande do Sul (UFRGS)
Rua Sarmento Leite 320, 90050-170, Porto Alegre–RS, Brazil
krafta@ufrgs.br

Abstract. This paper presents the MAS-SOC approach to Multi-Agent Based Simulation. It integrates specific agent technologies for agent programming and communication, and includes a language we have designed for the specification of the environment to be shared by the agents in a simulation. A graphical interface is provided which helps the development of agent simulations (by managing libraries of simulation components and automatically generating appropriate source codes for the associated interpreters). In future improvements of this approach, we aim at including extra features that would favour the development of social simulations in particular, and to further improve the user interface so as to facilitate the access of social scientists to the design and implementation of multi-agent based simulations. In order to assess our platform for agent simulation, a case study on social aspects of the production and occupation of urban spaces is under development; this paper also briefly describes that social simulation and its preliminary results.

1 Introduction

The main goal of the MAS-SOC project (**M**ulti-**A**gent **S**imulations for the **SOC**ial Sciences) is to provide a framework for the creation of agent-based simulations which does not require too much experience in programming from users. In particular, it should allow for the design and implementation of cognitive agents. A graphical user interface is provided which facilitates the specification of multi-agent environments, agents (their beliefs and plans), and multi-agent simulations; it also helps the management of libraries of these simulation components. From the information given by the user, the system generates source codes for the interpreters of the language for programming cognitive agents and the language for the specification of multi-agent environments on which MAS-SOC is based.

* Currently at the Department of Computer Science, University of Liverpool, U.K.: R.Bordini@csc.liv.ac.uk

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 70–91, 2004.

In our approach, the reasoning of agents is specified in AgentSpeak(XL), an extension to AgentSpeak(L) [31] that we introduced in [2]. The environments where agents are to be situated are specified in ELMS, a language we have designed for the description of multi-agent environments specifically. The development of an environment description language for our simulation platform was needed because when a multi-agent system is a (completely) computational system (i.e., not situated in the real world), this is an important level of the engineering of multi-agent systems. However, this level of agent-oriented software engineering is not normally addressed in the literature, as environments are simply considered as "given", in particular when cognitive agents are specifically targeted.

The interactions among the simulation components (i.e., agent-agent, agent-environment, and the graphical interface for creating and controlling the simulations) is implemented with the SACI toolkit [19]. However, we still lack the means for specifying social structures explicitly (e.g., groups, organisations), which is very important for social simulation; we discuss this in detail towards the end of the paper. To provide mechanisms for specifying such structures is part of our long term objectives, which also include an attempt to reconcile cognition and emergence. This latter objective is inspired by Castelfranchi's [4] idea that only social simulation with cognitive agents ("mind-based social simulations" as he calls it) will allow the study of agents' minds individually and the emerging collective actions, which co-evolve determining each other. In others words, we aim (in the long term) at providing the basic conditions for MAS-SOC to help in the study of a fundamental problem in the social sciences, which is of the greatest relevance in multi-agent systems as well: the micro-macro link problem [6].

This paper is structured as follows. The next section briefly describes the main features of AgentSpeak(XL), the language used in MAS-SOC to specify the high level reasoning of agents. Section 3 presents ELMS, the language for specifying multi-agent environments. The general functioning of MAS-SOC simulations is explained in Section 4. Section 5 briefly mentions the first case study carried out according to the MAS-SOC approach; it aims at giving examples (to the extend that space permits) of MAS-SOC's underlying agent technologies. For that case study, a social simulation related to social aspects of urban growth is being developed. Preliminary results of these simulations are also given in section 5. With this and other case studies we aim at assessing our platform for agent simulation and improve it based on practical experience. Finally, in Section 6 we discuss related work, before we conclude the paper, also mentioning our long-term objectives for the MAS-SOC project.

2 AgentSpeak(XL)

In [2], we proposed several extensions to AgentSpeak(L), a programming language for BDI agents defined by Rao [31]. This section presents the main characteristics of AgentSpeak(L), and at the end of the section we mention some of the extensions we proposed to it.

AgentSpeak(L) is a BDI agent-oriented programming language introduced in [31]. In that paper, not only has Rao defined formally the operation of an abstract interpreter for it, but he also sketched a proof theory for that language in which, he claimed, known properties that are satisfied by BDI systems using BDI Logics [30] could also be proved; further, he claimed that there is a one-to-one correspondence between his interpreter and the proof system. In this way, he proposed what can be considered the first viable approach to bridging the gap between BDI theory and practice, an important issue in autonomous agent research that has been widely discussed for a long time.

Further formalisation of the abstract interpreter and missing details were given by d'Inverno and Luck in [12]. Their formalisation was done using the Z formal specification language. In [27], a structural operational semantics for AgentSpeak(L) was given. A recent paper [1] introduced a way in which to define the informational, motivational, and deliberative modalities of BDI logics for AgentSpeak(L) agents, according to its operational semantics; this framework was then used to prove which of the Asymmetry Thesis Principles [30] apply to AgentSpeak(L) agents. This can be considered a step forward in bridging the gap between theory and practice of BDI systems. However, until recently there was no available implementation of AgentSpeak(L) based on Rao's abstract interpreter. In [26], a means for running AgentSpeak(L) programs within Sloman's SIM_AGENT framework [34] was described. That was the first prototype implementation of an AgentSpeak(L) interpreter; it was called SIM_Speak. A mechanism is provided in SIM_Speak for the conversion of AgentSpeak(L) programs into running code within SIM_AGENT. For the extended language AgentSpeak(XL) [2] (described later in this section), an efficient interpreter in C++ was implemented from scratch.

We now cover the basics of the syntax and informal semantics of AgentSpeak(L)(further details can be found in the references given above). An AgentSpeak(L) agent is created by the specification of a set of base beliefs and a set of plans. Base beliefs are ground atoms in the usual form (e.g., as in Prolog). The set of *beliefs* represents the information an agent has about the world (i.e., the environment and other agents). Plans are sequences of actions (or goals) an agent needs to execute (or achieve) in order to handle some perceived event.

AgentSpeak(L) distinguishes two types of goals: *achievement goals* and *test goals.* The former are predicates as defined for beliefs but they are prefixed with the '!' operator, while the latter are prefixed with the '?' operator. Achievement goals are used when the agent needs to achieve a certain state of the world (by performing actions and achieving subgoals, i.e., by executing a plan). Test goals are used when the agent needs to test whether the associated predicate is believed to be true, i.e., whether there is a unifying function which makes it a logical consequence of the agent's current belief base (thus further binding variables in the body of a plan instance).

Next, the notion of *triggering event* is introduced. It is a very important concept in this language, as an AgentSpeak(L) agent reacts to events by executing plans. Events happen as a consequence of changes in beliefs during the process of belief revision based on perception of the environment, or additions of goals due to the execution of plans. There are two types of triggering events: those related to the *addition* ('+') and those related to the *deletion* ('–') of mental attitudes, specifically beliefs and goals (e.g., $-busy(line)$, $+!book(tickets)$).

Plans need to refer to the basic *actions* that an agent is able to perform on its environment (so as to change it). An AgentSpeak(L) *plan* has a head which is formed of a triggering event (the purpose of that plan), and a conjunction of belief literals forming a context that needs be satisfied (the context must be a logical consequence of that agent's belief base) for the plan to be considered applicable at that moment for handling a particular event. A plan has also a body, which is a sequence of basic actions or (sub) goals that the agent has to achieve (or test).

An AgentSpeak(L) *agent* is given by a tuple $\langle E, B, P, I, A, \mathcal{S}_{\mathcal{E}}, \mathcal{S}_{\mathcal{O}}, \mathcal{S}_{\mathcal{I}} \rangle$, where E is a set of events, B is a set of base beliefs, P is a set of plans, I is a set of intentions, and A is a set of actions. The selection function $\mathcal{S}_{\mathcal{E}}$ selects one event from E (the one to be handled in a particular reasoning cycle); the selection function $\mathcal{S}_{\mathcal{O}}$ selects one plan (i.e., an option) from a set of applicable plans (for handling the chosen event); and $\mathcal{S}_{\mathcal{I}}$ selects an intention from I (the one that will be executed one step further at that reasoning cycle). The *selection functions* are supposed to be agent-specific; however, AgentSpeak(L) provides no means for specifying them, although they essential in the interpretation of an AgentSpeak(L) program.

Finally, *intentions* are particular courses of actions to which an agent has committed in order to handle certain events. Each intention is a stack of partially instantiated plans. Events, which may start off the execution of plans, can be external, when originating from perception of the agent's environment (more precisely, external events are addition and deletion of beliefs due to belief revision); or internal, when generated from the agent's own execution of a plan (e.g., a subgoal in a plan is an addition of goal which is triggering event for another plan). In the latter case, the event is accompanied of the intention which generated it (as the plan chosen for that event will be pushed on top of that existing intention). External events create new intentions in I, representing the various focuses of attention for the agent's action on the environment.

We now turn to AgentSpeak(XL), an extension we have proposed to AgentSpeak(L). In another project, we have been working towards the integration of cognitive and utilitarian (decision and game-theoretic) approaches to multi-agent systems. In [2], an initial contribution towards that direction was made. The idea was to use TÆMS [8] and the Design-To-Criteria (DTC) scheduler [35]—see [24] for an overview of that approach to multi-agent systems—to provide greater expressive power for BDI programming languages by addressing questions such as intention selection.

AgentSpeak(XL) extends AgentSpeak(L) for improving that language in various ways, such as handling plan failure, belief addition and deletion in the bodies of plans, communication, and provides a new construct called *internal actions* which allow for general extensibility of the language. In [2], we concentrated on the use of internal actions to accommodate the on-the-fly use of DTC for generating efficient intention selection functions. The extended language allows one to express relations between plans, as well as quantitative criteria for their execution. This has greatly improved the expressiveness of the language, facilitating the programming of certain types of applications where quantitative reasoning or priorities among certain tasks are required. In fact, it has provided programmers with control over an agent's intentions (i.e., an agent's multiple focuses of attention) which was not possible in the original AgentSpeak(L) interpreter (unless an agent-specific intention selection function was implemented with ordinary programming languages).

The AgentSpeak(XL) interpreter is available as free software[2]. This language is used as part of the MAS-SOC approach for the creation of the individual agents that participate in the simulations. A graphical interface helps the user in managing libraries of plans and agent definitions, and the selection of individual agent instances to participate in a simulation (this is discussed further in Section 4.2).

3 ELMS: An Environment Description Language for Multi-agent Simulations

Agents in a multi-agent system interact with the environment where they are situated and interact with each other (possibly through the share environment). Therefore, the environment has an important role in a multi-agent system, whether the environment is the Internet, the real world, or some virtual environment.

This section introduces the main aspects of the language we created for the specification of a simulated (i.e., virtual) environment that is to be shared by the agents in a multi-agent system. The language is called ELMS (**E**nvironment Description **L**anguage for **M**ulti-Agent **S**imulation).

3.1 Multi-agent Environments

Agents are computational systems situated in some environment, and are capable of autonomous actions in this environment in order to meet their objectives [37]. Agents perceive and interact with each other via the environment, and they act upon it so that it reaches a certain state where their goals are achieved. Therefore, environment modelling is an important issue in the development of multi-agent systems where the agents do not act directly on a physical or existing environment (e.g., as robots with real sensors and effectors, or Internet agents). This applies to reactive as well as cognitive agent societies (as discussed below). Nevertheless, the multi-agent systems literature seldom considers this part of the engineering of agent societies (as environments are assumed as given), in particular in association with cognitive agents.

In a reactive multi-agent systems, the environment plays a major role. Since reactive agents have no memory and no high-level (i.e., speech-act based) direct communication with other agents, it is only perception of the environment that allows them to make decisions on how to act. On the other hand, cognitive agents have an internal representation of the environment, yet they make decisions (e.g., to adopt new goals, to change courses of actions) based on the changes that perception of the environment causes on that representation. Thus, environment modelling is equally important for both classes of multi-agent systems. Although some multi-agent systems may be situated in an existing environment, in agent-based simulations, the environment is necessarily a computational process as well, so modelling multi-agent environments is always an important issue.

In [32], a number of characteristics that can be used to classify environments is given. We recall them briefly below, so as to characterise the classes of environments that can be defined with ELMS.

[2] URL `http://protem.inf.ufrgs.br/cucla/` then click “Downloads”.

Accessible vs. inaccessible: If the agent can perceive (through its sensors) all the relevant properties to its deliberation process, then this environment is said to be accessible to the agent.

Deterministic vs. nondeterministic: An environment is deterministic if its next state is completely determined by the current state and the actions performed by the agents. Of course, an environment can appear to be non-deterministic from the point of view of an agent if other agents perform actions on the environment and the environment is inaccessible to that agent.

Episodic vs. non-episodic: An environment is episodic if an agent's experiences are independent from one another. An action executed in one "episode" will not affect the next ones (where an episode consists of the agent perceiving the environment and acting accordingly).

Static vs. dynamic: An environment is said to be dynamic from the point of view of an agent if the environment can change during the agent's deliberation process. In a MAS where multiple agents perform simultaneous actions asynchronously, the environment is certainly dynamic. If the environment is static but the agent's performance score decreases with the passage time (during deliberation), then the environment is said *semi-dynamic.*

Discrete vs. continuous: An environment is said to be discrete if its attributes have a limited number of distinct possible values and well defined perceptions and actions.

With ELMS, it is possible to specify environments that are (from the point of view of the agents): inaccessible, non-deterministic, non-episodic, and dynamic; however, they have to be discrete. This class of environments is the most complex and comprehensive, except for the class of environments that are continuous beside all that. However, continuous environments are notoriously difficult to simulate. Although it is not possible to define continuous environments in ELMS, we believe that it allows the definition of rather complex environments, supporting a wide range of multi-agent applications (in particular for social simulation).

3.2 The MAS-SOC Approach to Environment Modelling

An environment description is a specification of the properties and behaviour of the environment. In our approach, such specification includes mainly sets of: objects, to which we interchangeably refer as *resources* of the environment; agents, or more precisely, their "physical" representation that is visible to other agents in the environment); actions that each type of agent can perform in the environment; reactions that object displays when agent actions affect them; the perception levels available to each type of agent; and the properties to which external observers (e.g., the users) have access.

The resources (i.e., objects) that are present in an environment can be modelled as a set of properties and the actions that they can perform in response to stimuli (that are external to the object). That is, objects can *react*—only agents are pro-active. Agents can be considered components of the environment insofar as, from the point of view of one agent, any other agent are special components of the environment (only certain properties of an agent can be perceived by other agents, and this must be specified by designers of agent-based simulations). Thus, in the environment description, agents are defined by a list of properties (which defines the perceptible aspects of agents), a list of actions that they are able to execute (pro-actively), and a list of perception levels to which they have access. From the point of view of the environment, the deliberation activities of an agent are not relevant, since they are internal to the agent, i.e., not observable to the other agents. As mentioned before, the internal aspects of agents are described with the use of AgentSpeak(XL), as seen in Section 2.

Quite frequently, spatial aspects of the environment are modelled in agent simulations by means of a grid. ELMS provides a number of features for dealing with grids, if the designer of the environment chooses to have one. In the constructs that make reference to the grid, positions can be accessed by absolute or relative coordinates. Relative coordinates are prefixed by '+' and ' –' signs, so (+1, –1, +0), for example, refers to the position on the upper right diagonal from the agent's present position.

For the definition of the perception levels to which each type of agent has access, it is necessary to define which properties of the environment, agents, and objects are to be perceived at each level (i.e., type) of perception. The conditions associated with each perceptible property can be specified as well; that is, users can state that the specified properties will be informed to agents with that perception level, when perception is to be sent to them, only under certain conditions. An action is defined as a sequence of changes in properties (of the environment, resources, or agents) that it causes, and the preconditions that must be satisfied for the action to be executed in the environment at all.

3.3 Language Constructs in ELMS

The ELMS language uses an XML syntax, which can be a somewhat cumbersome to be used directly; however, recall that the specification of environments is to be done through the MAS-SOC graphical interface (see Section 4.2). Nevertheless, environment specifications can be written directly in XML with a simple text editor, or some other tool, if the user prefers to do so.

An environment specification in ELMS can be formed of nine types of definitions. There are specific language constructs for each of these definitions, which can be repeated any number of times, although in principle not in any arbitrary order[3]. The usual order for the definitions is as follows:

Grid Options: The grid definition is optional. A grid can be two-dimensional or three-dimensional, the parameters being the sizes of the X, Y, and Z axes, and there is a wrap-around option. Still within the grid definition, a list of cell attributes can be given; the attributes defined there will be replicated for each cell of the grid. The attribute definition comprises its name, the type of the property (integer, float, string or boolean), and an optional initial value.

Resources: In a resource definition section, the classes of resources (or objects) are defined; later, during simulation, several instances of such classes may be allocated.

[3] Recall that the definitions are entered in any order in a graphical interface; the sequence of definitions in the appropriate order is generated automatically. However, there is no explicit syntactic division of these definition sections in the source file.

A definition of a resource class includes the class name, a list of attributes and a set of reactions. The attributes are defined in the same way as the cell attributes (i.e., through the specification of a name, type and initial value). The reactions that a class of resources can have is given by a list of the names (i.e., labels) identifying those reactions (see below how reactions are defined).

Agents: In this part of an environment specification, we find the definitions of the classes of agents that may participate in a simulation with that environment. A specification of an agent class contains its name, a list of attributes, a list of actions, and a list of perception levels. The list of attributes is defined as before; it characterises the observable properties of agents, from the point of view of the environment and other agents. It is then necessary to specify a list of the action names that agents of that class are allowed to perform in that environment. The set of perceptions is a list of the names of perception levels (see below) that are available to that class of agents (i.e., the information that the environment will send to participating agents of that class at every reasoning cycle). Note that the same perception and action names can appear in any number of definitions, that is, they can be reused in different classes of agents (and equally with reactions for resources).

Perceptions: This construct allows the specification of the perception levels listed in agent specifications. A perception level definition is formed by a name, an optional list of preconditions, and a list of properties names that are perceptible. The listed properties can be any of those associated with the definitions of resources, agents, cells of the grid, or simulation control variables. If all the preconditions (e.g., whether the agent is located on a specific position of the grid) are all satisfied, then the values of those attributes (properties) will be sent to the agent as the result of its perception of the environment. Note that perception can be based on the spatial position of the agent, but this is not mandatory; any type of perception can be defined by the designer of the environment.

Actions: In this section of the environment description, the actions that appeared in agent definitions are described. An action definition includes its label, an optional list of parameters, an optional list of preconditions, and a sequence of commands which determine what changes in the environment the action causes. The list of parameters tells what parameters will be received from the agent for the execution of that type of action. The possible commands defining a action are assignments of values to attributes, and allocations or repositioning of instances of agents or resources within the grid. Resources can also be instantiated or removed by commands in an action. If the preconditions are all satisfied, then all the commands in the sequence of commands will be executed, changing the environment accordingly[4].

Reactions: This part of the specification is where the possible reactions of the resources in the environment are defined. For each reaction, its name, a list of preconditions, and a sequence of commands is given. The commands are exactly as for actions (seen

[4] Since agents are constantly perceiving, reasoning and acting, the actions they execute in the environment should normally atomic. That is, it is known before the next reasoning cycle whether the action was successfully executed, and if it was, its perceptible effects will be noticed by the agent when it does belief revision just before the next reasoning cycle. Although it is possible to make alternative design choices where actions are not atomic, it seems that simulations should be more appropriately engineered that way.

above). All expressions in the list of preconditions must be satisfied for the respective reaction to take place. Differently from actions, where only one action (per agent) is performed, all reactions that satisfy their preconditions will be executed in that single simulation cycle.

Observables: This is where the user defines which properties of the agents, resources and the environment itself will be sent to the MAS-SOC interface as the result of a simulation cycle. The properties to be selected as observable can be any of those associated with instances of resources and agents, cells of the grid, and simulation control variables.

Simulation Values: This section defines the current values of the environment control variables and the values of the properties of each instance of agents and resources. Also in this section, the position of the agents and resources in the grid are defined (if that is the case). The values for environment control variables can be defined by assignment commands over predefined variable names. The values for agent and resource properties are accessed by an agent (or resource) class name, an index (of a particular instance), and then the property name. The positions of agents and resources in the grid are determined by a language construct which requires a cell position and the list of resources and agents that are located on that grid cell. This section is particularly useful in specifying simulation snapshots (which are mentioned in Section 4.1).

Initialisation: Finally, this part of the specification allows resources in the environment to be instantiated and allocated to grid positions in the initial state of the simulation (resources can also be created dynamically in the environment). All commands in this section are only executed before the start of the simulation.

Note that ELMS allows quite flexible environment definitions. It is up to the environment designer to decide what aspects of the environment can be perceptible to agents and observable to users (as well as defining how actions change the environment). Such aspects can be modelled as properties of resources, agents, and even actions if they change some resource or agent properties accordingly.

4 MAS-SOC Simulations

This section describes how ELMS environment specifications are executed, and also shows how all MAS-SOC simulation components interact.

4.1 Running ELMS Environments

In MAS-SOC, the environment execution is controlled by a process which, in each simulation cycle, sends to all agents in a simulation the perceptions to which they have access (as specified in ELMS). Perceptions are transmitted in messages as a list of AgentSpeak(L) ground atoms. After sending the perceptions, the process waits for the actions that the agents have chosen to perform in that simulation cycle (after they executed an internal reasoning cycle). This process, called *environment controller,* is automatically generated from the environment specification written in the language mentioned in the previous section.

The main responsibilities of the environment controller process are to:

1. execute the commands in the initialisation section before the start of the simulation;
2. check which perceptions from the agent's perception list are in fact available at that time (i.e., check which of an agent's perceptible properties satisfy the specified conditions);
3. send the resulting perceptions (those that satisfied the conditions) to the agents;
4. receive from all agents the actions that they have chosen to perform in that cycle;
5. randomise the action queue, to assure that each agent has a chance to execute its action first;
6. check which actions, from the ones received from the agents, satisfy the respective conditions for execution;
7. execute the actions that were determined as enabled for execution;
8. send the values of the observable properties to the interface;
9. maintain an internal record of the agents that enter and leave the society.

Users can define whether the simulation will be run in synchronous or asynchronous mode. Running a simulation in synchronous mode means that the environment wait for all agents to choose an action to perform; only then all actions are executed, and after that perception is sent to all agents. Item 5 refers to a queue randomisation that is performed to ensure that all agents have, on average, similar chance of executing their action first when synchronous simulation is being used. As MAS-SOC simulations may be distributed, that mechanism guarantees that network or processor performances will not interfere with the results of synchronous simulations.

As mentioned earlier, the environment controller process is responsible for the actual running of an environment; it updates and controls the access to the data structures that represent that environment. The data structures that represent the environment are generated by the ELMS interpreter for a specification passed on to it as input. The language has constructs that allow the use of an ELMS specification as a snapshot of a simulation. The environment controller process can generate such snapshots from the data structures. This feature allows users to save a simulation status for later execution, or to make on-the-fly changes in the environment (via the interface, or changing the ELMS text manually). This should also be useful in future work on providing complex forms of visualisation of multi-agent simulations.

MAS-SOC uses the SACI toolkit [19] for implementing the communication among the agents, the environment, and the interface that constitute any particular simulation. SACI supports KQML-based communication, and provides an infrastructure for managing distributed agents. All agents participating in a simulation, the interface, and the environment controller are registered to a SACI society. Through SACI, every member of the society can communicate to others members of that society by simply sending messages addressed to that member's name within the society (making transparent all issues of the distributed operation of the simulation components). Therefore, it should be possible for any SACI-based agent to interact with the other simulation components. An interesting use of this feature, for example, would be in providing an interface for human agents to take part in a simulated society (although this is not currently one of the main goals of the MAS-SOC project). This feature of open SACI societies can also be very useful in debugging and analysing simulations (by the introduction of "observer" agents).

SACI is available as free software[5]. The ELMS interpreter too will be made available as free software soon, as will the whole MAS-SOC platform[6] eventually (when sufficiently tested on practical applications).

4.2 Creating Simulations with MAS-SOC

A graphical user interface which facilitates the creation and running of simulations in being developed. This interface gives access to what we call the MAS-SOC *manager*. Besides facilitating the creation of simulations, the MAS-SOC manager integrates the various technologies used in our approach, such as the ELMS interpreter, the AgentSpeak(XL) interpreter, and the SACI infrastructure.

The first aspect of the MAS-SOC manager that should be mentioned is related to the creation of libraries of plans, agents and environment (maintained in separate files), which facilitates the reuse of those definitions in different simulations. The MAS-SOC manager allows the creation and edition of each of these libraries. A file containing a plan library consists of a series of plan definitions in the AgentSpeak(XL) syntax. In an agent library, each agent[7] is represented by a name, a set of AgentSpeak(XL) base beliefs (the initial beliefs that agents of this type will have when the simulation begins), and a list of pointers to plans (in fact, plan labels) in specific plan libraries. With this, the AgentSpeak(XL) source codes for the agents can be generated by the interface and sent to the running instances of the AgentSpeak(XL) interpreter. The information for an environment definition is also prompted from the graphical interface, and the MAS-SOC manager automatically generates the XML-based ELMS source code, which is sent to the ELMS interpreter.

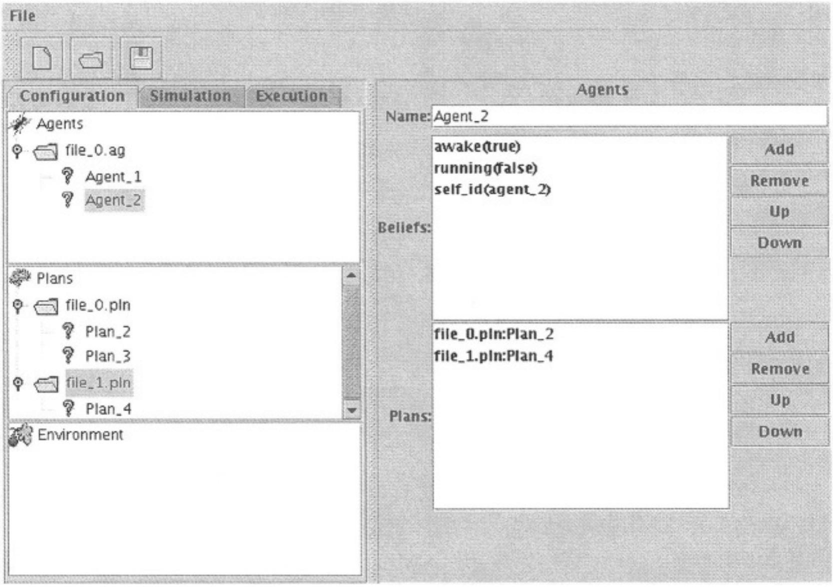
Figure 1 gives a flavour of the MAS-SOC user interface. It has the style of a "workspace", where one can create and edit libraries of plans, agents, and environments, which can then be used in defining a multi-agent simulation. Plans and agents follow the straightforward syntax of AgentSpeak(XL), and all necessary information for an ELMS environment description if prompted via a form-like interface. Other feature of the MAS-SOC manager are the creation, execution and monitoring of the simulation, which we are still improving. This part of the platform provides the integration of the several systems forming the MAS-SOC approach.

Figure 2 gives an overview of the functioning of the various parts of MAS-SOC that are controlled by the manager, and shows how they relate to each other. Through the "GUI" (Graphic User Interface) of the "MAS-SOC MANAGER", the user defines the agents and the MAS-SOC manager generates the appropriate "ASPK Code", which means an agent definition in the AgentSpeak(XL) language. Also defined through the GUI is the "ELMS Code", an environment description in the ELMS language. After the environment and agent codes have been prepared, the MAS-SOC manager starts

[5] URL `<http://www.lti.pcs.usp.br/saci>`.

[6] URL `<http://www.inf.ufrgs.br/~massoc>`.

[7] In fact, this refers to *types* of agents, as each of these agent definitions may have various instances in a simulation.



Name: Agent_2

Beliefs:
awake(true)
running(false)
self_id(agent_2)

Plans:
file_0.pln:Plan_2
file_1.pln:Plan_4

Fig. 1. The MAS-SOC Manager GUI.

the "Environment Controller". In its turn, the environment controller reads the ELMS specification, generating the appropriate data structures representing that environment. After that, the SACI society is created, through which the agents, the environment, and the MAS-SOC manager communicate. Then the agents are created by running instances of the AgentSpeak(XL) interpreter, each receiving the AgentSpeak(XL) code for one of the agents, as defined by the user. The agents connect themselves to the SACI society, so that through it they will receive the relevant perception of the environment and will send the actions they have chosen to perform so as to change the environment.

In a simulation definition window of the user interface, the user determines the set of individual agents and the particular environment that are intended for a given simulation. From the environment definition, the MAS-SOC manager checks which types of agents can participate in the simulation, and allows the user to choose, for each of those types, the number of instances of individual agents that will be created (by means of the SACI toolkit). Each of these agents runs an AgentSpeak(XL) interpreter with the source code generated by the MAS-SOC manager. After the user has informed the chosen environment and the instances of agents, the simulation can be started off. The execution of agents can aborted and new ones can be created through the MAS-SOC manager.

An execution window then provides the information about the components of a simulation (agents and resources) which are active in a simulation. There are in fact two levels of information about agents: their internal and external states. The internal state gives information on an agent's mental attitudes (e.g., its present beliefs and intentions) at each simulation step, while its external state is related to the characteristics (properties)

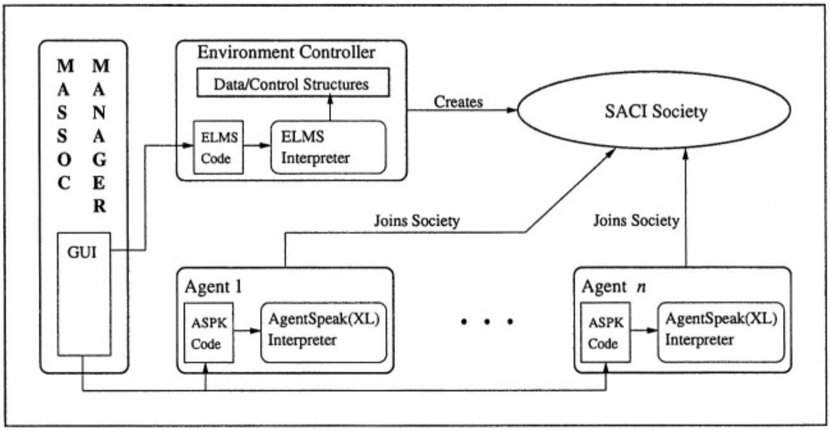


Fig.2. Creating a MAS-SOC Simulation.

of the agent that are perceptible to other agents through the environment. These two levels of information on agents can be accessed separately from the execution window. For the resources, one can observe the current values associated with their properties (attributes).

We now explain in detail all that happens before a simulation starts and how the whole simulation is run (some of these steps are depicted in Figure 3):

1. the MAS-SOC manager starts up the ELMS interpreter, which processes the environment specification given as input;
2. the interpreter generates the data structures that will be used to simulate the specified environment;
3. the initialisation section of the environment is executed;
4. the environment controller process (generated by the ELMS interpreter) creates a SACI society and registers itself as a member of the society;
5. the MAS-SOC manager registers itself as member of the SACI society;
6. for each agent in the simulation, the manager creates a process running the AgentSpeak(XL) interpreter (giving the appropriate source code as input);
7. all agents register themselves as members of the SACI society;
8. the manager checks whether everything is ready for the simulation to start (i.e., checks whether all agents and the environment controller are running);
9. the manager sends a "step" or "run" signal to the environment controller (according to the user's command);
10. the environment sends one round of perceptions to the agents;
11. agents run one reasoning cycle (after the belief revision process based on the perceptions received) and each agent informs to the environment which action it has chosen to perform in that cycle;
12. the environment executes the actions received from the agents;
13. the environment executes the reactions of the resources to the agent actions;
14. the environment sends the values of the observable properties to the interface;

and the cycle is now repeated from step 9, until the user chooses to stop the simulation (if a “run” command rather than “step” was issued).

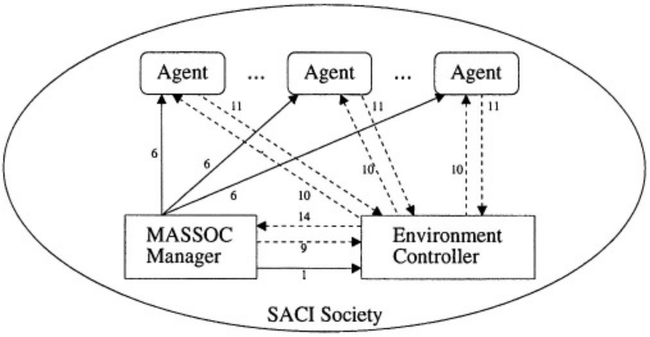


Fig. 3. MAS-SOC Components and Simulation Cycle.

Figure 3 shows the structure of a MAS-SOC simulation and the main interactions forming a simulation cycle. The “Agent” boxes represent AgentSpeak(XL) interpreters running the agent codes defined with the help of the MAS-SOC manager. The “Environment Controller” box represents the running process based on the data structures generated by the ELMS interpreter and additional structures to control the execution of environments. The “SACI Society” title represents the fact all interacting components of a simulation are SACI agents (to allow for their creation and communication over a computer network).

This concludes the overview of the functioning of a MAS-SOC simulation. We next mention a case study in which we have been working in order to evaluate and improve our approach to multi-agent based simulation.

5 A Case Study on the Simulation of Urban Growth

We are in the process of implementing a MAS-SOC simulation related to social aspects of urban growth. This and other simulations will be conducted using MAS-SOC for assessing our approach and improving the platform. In this section, the urban growth simulation is briefly presented as illustration of the approach to multi-agent simulation we propose in this paper.

This social simulation of urban growth includes three types of agents: consumers of commercial space, consumers of residential space, and developers (who build and sell properties). The city (i.e, the environment where agents are situated) in this application is formed by Built Form Units (BFUs), representing properties. Each BFU is formed by a number of basic plots, depending on the size and use of the property it represents. The plots all have the same size, and form the whole territory of the city, as a square grid. The price of the properties depends on factors such as land value, the type and age of the built units and the status of the neighbourhood (i.e., the social classes of the agents occupying the properties around). Our main objectives with this simulation are, first, to investigate how certain relations among social classes regulate urban growth and their impact on the spatial form of cities, and second, to provide better understanding of the kinds of interaction between social agents and space, which causes spatial macro orders and social behaviour to emerge [28,22].

In this simulation, the agents are situated in an environment which models a city. These agents interact through the execution of certain actions that they can perform in the environment, as described next. The developers can build six different types of constructions: three types of residential buildings (A, B, or C, relating to social classes[8]) and three types of commercial buildings (Small, Medium, or Large). Consumers of residential space can buy, from developers, residential properties of the appropriate type for their social class. When choosing properties to buy, agents also take into consideration the social class of the agents living in the neighbourhood of the properties: agents of the upper social classes avoid living in neighbourhoods of the lower classes, whereas middle-class agents attempt to buy properties in upper-class neighbourhoods. Consumers of commercial space can buy, from developers, commercial buildings of the appropriate type for their number of clients.

Those three types of agents were implemented in AgentSpeak(XL) and, as an illustration, we show in Figure 4 a preliminary implementation of the simplest type of agent (the commercial consumers).

Beliefs:

```
id(ms1).
size(big).
occupy(1).
service(music_shop).
money(1000).
step(20).
```

Plans:

```
+step(N) : size(B) & service(S) & money(M)
    ← find_commercial(B,S,M);
+offer(EA,BFU) : ug.worthy(BFU)
    ← ?id(Self);
      .send(EA,request,buy(BFU,Self));
+sell(EA,BFU) : true
    ← occupy(BFU).
```

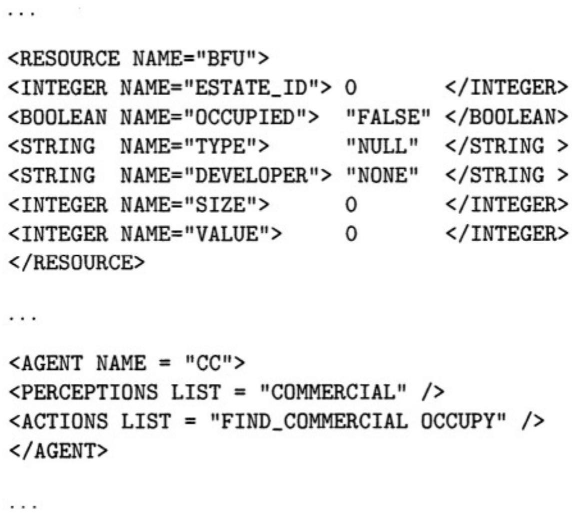
Fig.4. AgentSpeak(XL) Code for Commercial Consumer Agents.

In this agent, `id` and `size` are examples of *initial* beliefs. As usual with AgentSpeak(XL) agents, beliefs are created and changed during the simulation. This very simple agent having only three plans is sufficient to specify an agent having the role of commercial consumer in the system. The `.send` action is a special one which allows inter-agent communication with particular illocutionary forces (e.g., `request`). The internal action[9] `ug.worthy(U)` is a C++ function which calculates whether a BFU offered by a developer agent is worth buying (from the point of view of consumer agents).

[8] In our simulation, agents of type "consumers of residential space" belong to social classes, which is reminiscent of Hales's simulations of tag-based groups of agents [17].

[9] The construct of *internal actions* is part of the extension to AgentSpeak(L) we presented in [2], see Section 2. The `ug` library of internal actions was implemented to provide some specific functionalities we needed for this urban growth simulation.

The environment (i.e., the city) has been specified in ELMS and with this case study we were able to identify key features and difficulties of the ELMS/MAS-SOC approach; ELMS in particular was improved substantially based on the work on this case study. As a brief example, Figure 5 shows a small sample of the ELMS code for the environment in this simulation. Attributes may have their initial values set in the environment definition, and can be changed during the simulation. The part of the environment definition shown in the figure covers only the definitions of the "BFU" resource and the "commercial consumer" agent.



```
...

<RESOURCE NAME="BFU">
<INTEGER NAME="ESTATE_ID"> 0       </INTEGER>
<BOOLEAN NAME="OCCUPIED">  "FALSE" </BOOLEAN>
<STRING  NAME="TYPE">      "NULL"  </STRING >
<STRING  NAME="DEVELOPER"> "NONE"  </STRING >
<INTEGER NAME="SIZE">      0       </INTEGER>
<INTEGER NAME="VALUE">     0       </INTEGER>
</RESOURCE>

...

<AGENT NAME = "CC">
<PERCEPTIONS LIST = "COMMERCIAL" />
<ACTIONS LIST = "FIND_COMMERCIAL OCCUPY" />
</AGENT>

...
```

Fig. 5. Sample of the ELMS Specification for the City Environment.

A very important type of resource in this environment is the BFU (i.e., the built forms that agents can sell and occupy). For this reason, we have used the BFU resource definition in the example of the ELMS specification. Each of the BFU attributes is briefly explained below:

- ESTATE_ID: this is the identification of the property represented by this BFU;
- OCCUPIED: when "TRUE", indicates that the BFU has been occupied by an agent;
- TYPE: indicates the type of building represented by the BFU;
- DEVELOPER: the identification of the developer that has built the BFU;
- SIZE: registers the size of the BFU in terms of the number of basic plots (i.e., territory units) forming it;
- VALUE: the financial value of the BFU in the current simulation cycle.

The definition of commercial consumer agents says that only percepts at the level labelled "COMMERCIAL" is sent by the environment to the commercial consumer agents. The actions "FIND_COMMERCIAL" and "OCCUPY" are the ones that this type of agent can perform in the environment.

A complete definition of the environment involves the use of other language constructs in ELMS to specify in details those actions, the perception level, as well as the other agents and resources. We included this short part of the specification as an example; a complete presentation of this application and ELMS itself should be given in future papers. The sample above uses the XML syntax of ELMS. Note, however, that users are not expected to type their specification in this rather clumsy notation. The MAS-SOC manager generates the ELMS source file automatically from the information given by the user through the graphical interface.

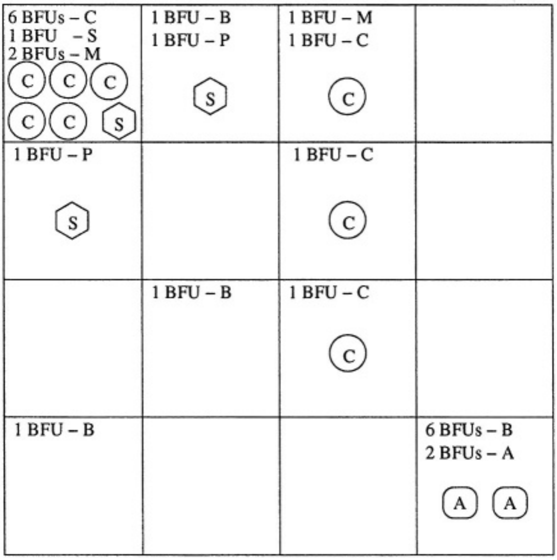


Fig. 6. Preliminary Simulation Results.

The simulation described in this section is still under development, but initial results can be seen in Figure 6. It shows one sample snapshot of a simulation with 55 agents of which 10 are developers, 28 are consumers of residential space (16 class C, 8 classB, and 4class A), and 17 are consumers of commercial space (10 type S, 5 type M, and 2 type L). What is shown in the figure is the state of the city after 400 simulation steps. Although we only have preliminary versions of the simulation specifications, and the simulation parameters have not been tuned properly yet, the simulation already shows interesting results where some class segregation can be observed; the proximity of services when relevant can also be observed.

6 Related Work

As mentioned in the introduction, at this stage we still lack a general mechanism for specifying social structures within our society of cognitive agents. However, to the best of our knowledge, there is no other implemented platform for developing BDI agents that also provides such mechanisms. Frameworks for team plans and social roles within the BDI architecture have been studied (e.g., in [5]), but not incorporated into working implementations. Numerous papers have appeared recently which propose various approaches for designing multi-agent organisations. However, none of them propose clearly an integrated mechanism for the implementation of cognitive agents, as we do here; all these platforms are conceived for reactive agents. For some interesting recent ideas on organisations in agent societies, see e.g. [20]. This is one likely source of inspiration for the work of introducing that level of abstraction in our approach (other well known sources of work on organisations are discussed below). Also in this volume one can find interesting papers on agent organisations, e.g. [10]. What is particular interesting in that paper is the logic for contract representation on which the authors are working; such logic would figure well among the many modal logics of interest to multi-agent systems. The ability to represent and enforce norms in agent organisations is also paramount; in this respect, two papers of particular interest which can also be found in this volume are [11] and [25].

MadKit [15,16] is a platform for building multi-agent systems based on an organisational approach. It is rooted on an "agent-group-role" model. Agent societies are built regardless of the agent architecture used for implementing them; however, in our opinion, the issue of integrating cognitive autonomous agents with such societal structures is not one that can be taken for granted. In fact, in MadKit the agents are assumed as given, it only provides templates for building reactive agents (as many other platforms do). The main difference of our approach to MadKit[10] regards the basic components of a multi-agent system on which the approaches focus. According to [9], the basic components of multi-agent systems are: agents, environments, interactions, and organisations. The MadKit platform focuses mainly on the organisational component and on interaction models. Our approach, on the other hand, focuses on environments and agents (in particular cognitive agents). In our approach, agents use the AgentSpeak(XL) language, following the widely studied BDI model of rational agency (even though the general structure of MAS-SOC allows the use of any type of agent, as long as it uses the SACI toolkit for communication and accepts the simple format of perceptions and actions we use for AgentSpeak(XL) agents).

Another source of ideas for the organisation level of multi-agent systems is the Gaia methodology [36]. As MadKit, Gaia is also based on organisational models aimed at the design of multi-agent systems that can be composed of heterogeneous agent models and theories. However, Gaia is a methodology for the design and analysis of agent-oriented system; it is not a platform for the development of multi-agent systems as MadKit is. Being a methodology for engineering multi-agent systems in general, it does not address important aspects of simulations, such as environment modelling, as we do in MAS-SOC. As with MadKit, again in Gaia there is not much emphasis on the engineering of individual cognitive agents that would work as part of the organisation.

Summarising the comparison with other approaches, in our approach environments are explicitly defined using the ELMS language, while in the other approaches, no provision is made for the specification of environments. The environment is either the "real world", or is simply assumed as "given". Also, while most other approaches focus on the organisational component and interaction structures of very simple (reactive) agents, the MAS-SOC approach focuses on the agents, environment and the agent-agent and agent-environment interactions. In particular, it allows for the easy implementation of *cognitive* agents, which is not normally the case of other platforms used for social simulation.

Specifying the social structure of an agent society is important for social simulation, but allowing the use of cognitive agents is also very important [4]. Improving our ap-

[10] We are using MadKit here as a representative of the various platforms for multi-agent organisations that are currently being used in social simulation. The comparison applies to most of them.

proach to cognitive agents so that it includes organisations is a more likely way forward in social simulation than the other current approaches which deal with organisations but provide no clear mechanism for building cognitive agents. At present, however, interaction structures are defined implicitly in each agent; there is no rigid exogenous interaction structures such as roles in our approach as yet. As a long term objective, we aim that our platform will allow for the conduction of simulations where organisations can emerge from the interactions among the agents, considering phenomena such as *immergence* and *second level emergence* [3,4]. We are particularly interested in Castelfranchi's point on reconciling cognition and emergence. This would bring multitudinous new possibilities to social simulations based on evolutionary approaches, as [17] in this volume, for example, without neglecting the cognitive aspect of agents.

Considering that an interpreter for an agent-oriented programming language is part of our approach, it is important to compare it with other agent-oriented languages as well. Since Shoham's paper on agent-oriented programming [33], many agent programming languages have been proposed, following various approaches. ConGolog [7] is a concurrent programming language based on the situation calculus, Concurrent MetateM [13] is based on temporal logics, and $\mathcal{MINERVA}$ [23] is based on dynamic logic programming. AgentSpeak(L) [31] is based on the BDI (Beliefs-Desires-Intentions) architecture [29] and the more practical experience with PRS [14] and dMARS [21]. Other BDI programming languages were derived from AgentSpeak(L), such as 3APL [18], improving it in certain ways (e.g., in handling plan failure). For our purposes, we found that extending AgentSpeak(L) was the best way forward. Its basic structure is simple, which is important for making the specifications of agents in MAS-SOC easy (hopefully, in the future, so simple that social scientists themselves can use it), and it has a neat notation. Also, it is more faithful, so to speak, in relation to the BDI architecture when compared to other BDI-oriented languages. The BDI architecture now permeates a significant part of research in autonomous agents and multi-agent systems, that is why it is important to consider BDI-based languages for our purpose of specifying cognitive agents.

7 Conclusion

We have presented a distinct combination of techniques from the area of multi-agent systems which we consider as most adequate for the construction of multi-agent based simulation, in particular social simulation. They are integrated within our MAS-SOC platform, which has a graphical interface for helping the management of libraries of plans, agents, environments and multi-agent simulations, as well as helping the control over running simulation. MAS-SOC is being implemented in JAVA, and although this is ongoing work, we have already experimented with it in an application in the area of urban growth. This initial application is helping in the process of improving the interface and the languages we are using, as well as the integration of the mentioned technologies. Preliminary results show that the approach is quite promising for the development of social simulations with many cognitive agents. Only the wide use of MAS-SOC (specially by social scientists) and comparison with other approaches can confirm our expectation that this is an adequate approach to agent-based simulation.

As future work, there are several improvements that we plan to do in our platform. In particular, we plan to concentrate on aspects of agent-based simulations which are particularly important for social simulation, such as the specification of social structures within agent societies, as discussed in the previous section, and the integration with free software packages for the statistical analysis of observed simulation results. In the long term, we aim at investigating the necessary mechanisms for reconciling cognition and emergence following the ideas in [4], and incorporating such mechanisms into MAS-SOC, thus allowing it to be used in investigations of the micro-macro link problem. Future work also include the progress with the simulation of social aspects of urban growth, and we are considering the implementation of various other social simulations.

Acknowledgements. This work was partially supported by CNPq and FAPERGS. We thank Marcelo G. de Azambuja for the period he joined the team of developers of the MAS-SOC project. The first author would like to thank Rodrigo Machado for his participation in the early stages of this project, and Cristiano Castelfranchi for sending a draft of his paper on the theory of social functions which, together with a tutorial on Social Simulation given by Edmund Chattoe, inspired the whole project.

References

1. Bordini, R. H. and Moreira, Á. F. 2002. Proving the asymmetry thesis principles for a BDI agent-oriented programming language. In Dix, J., Leite, J. A. and Satoh, K., eds., *Computational Logic in Multi-Agent Systems: 3rd International Workshop, CLIMA '02, Copenhagen, Denmark, August 1, 2002, Proceedings,* number 93 in Datalogiske Skrifter (Writings on Computer Science), 94–108. Roskilde University, Denmark.
2. Bordini, R. H., Bazzan, A. L. C., Jannone, R. O., Basso, D. M., Vicari, R. M. and Lesser, V. R. 2002. AgentSpeak(XL): Efficient intention selection in BDI agents via decision-theoretic task scheduling. In Castelfranchi, C. and Johnson, W. L., eds., *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-2002), 15–19 July, Bologna, Italy,* 1294–1302. New York, NY: ACM Press.
3. Castelfranchi, C. 1998. Simulating with cognitive agents: The importance of cognitive emergence. In Sichman, J. S., Conte, R. and Gilbert, N., eds., *Multi-Agent Systems and Agent-Based Simulation,* number 1534 in Lecture Notes in Artificial Intelligence, 26–44. Berlin: Springer-Verlag.
4. Castelfranchi, C. 2001. The theory of social functions: Challenges for computational social science and multi-agent learning. *Cognitive Systems Research* 2(1):5–38.
5. Cavedon, L. and Sonenberg, L. 1998. On social commitment, roles and preferred goals. In Demazeau, Y., ed., *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS'98), Agents'World, 4–7 July, Paris,* 80–87. Washington: IEEE Computer Society Press.
6. Conte, R. and Castelfranchi, C. 1995. *Cognitive and Social Action.* London: UCL Press.
7. de Giacomo, G., Lespérance, Y. and Levesque, H. J. 2000. ConGolog: A concurrent programming language based on the situation calculus. *Artificial Intelligence* 121:109–169.
8. Decker, K. S. and Lesser, V. R. 1993. Quantitative modeling of complex environments. *International Journal of Intelligent Systems in Accounting, Finance and Management* 2(4):215–234.

9. Demazeau, Y. 1995. From cognitive interactions to collective behaviour in agent-based systems. In *Proceedings of the European Conference on Cognitive Science.* Saint-Malo, April, 1995.
10. Dignum, V., Meyer, J.-J., Wiegand, H. and Dignum, F. 2002. An organisational-oriented model for agent societies. (In *RASTA 02 Pre-Proceedings, Hamburg University, Faculty of Informatics, Communications Vol.318).*
11. Dignum, F. 2002. Abstract norms and electronic institutions. (In *RASTA 02 Pre-Proceedings, Hamburg University, Faculty of Informatics, communications Vol.318).*
12. d'Inverno, M. and Luck, M. 1998. Engineering AgentSpeak(L): A formal computational model. *Journal of Logic and Computation* 8(3):1–27.
13. Fisher, M. 1994. A survey of concurrent METATEM—the language and its applications. In Gabbay, D. M. and Ohlbach, H. J., eds., *Temporal Logics—Proceedings of the First International Conference,* number 827 in Lecture Notes in Artificial Intelligence. Berlin: Springer-Verlag. 480–505.
14. Georgeff, M. P. and Lansky, A. L. 1987. Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI'87), 13–17 July, 1987, Seattle, WA,* 677–682. Manlo Park, CA: AAAI Press / MIT Press.
15. Gutknecht, O. and Ferber, J. 2000. The MadKit agent platform architecture. In *Agents Workshop on Infrastructure for Multi-Agent Systems,* 48–55.
16. Gutknecht, O., Ferber, J. and Michel, F. 2001. Integrating tools and infrastructures for generic multi-agent systems. In Müller, J. P., Andre, E., Sen, S. and Frasson, C., eds., *Proceedings of the Fifth International Conference on Autonomous Agents,* 441—448. Montreal, Canada: ACM Press.
17. Hales, D. 2002. The evolution of specialization in groups. (In this volume).
18. Hindriks, K. V, de Boer, F. S., van der Hoek, W. and Meyer, J.-J. C. 1999. Control structures of rule-based agent languages. In Müller, J. P., Singh, M. P. and Rao, A. S., eds., *Intelligent Agents V—Proceedings of the Fifth International Workshop on Agent Theories, Architectures, and Languages (ATAL-98), held as part of the Agents'World, Paris, 4–7 July, 1998,* number 1555 in Lecture Notes in Artificial Intelligence, 381–396. Heidelberg: Springer-Verlag.
19. Hübner, J. F. and Sichman, J. S. 2000. SACI: Uma ferramenta para implementação e monitorção, da comunicação entre agentes. In Monard, M. C. and Sichman, J. S., eds., *Proceedings of the International Joint Conference, 7th Ibero-American Conference on AI, 15th Brazilian Symposium on AI (IBERAMIA/SBIA 2000, Open Discussion Track), November 19–22, Atibaia, Sao Paulo, Brazil,* 47–56. São Carlos: ICMC/USP. <http://www.lti.pcs.usp.br/saci>.
20. Hübner, J. F., Sichman, J. S. and Boissier, O. 2002. $\mathcal{M}$OISE$^+$: Towards a structural, functional, and deontic model for MAS organization. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'2002), Bologna, Italy.* Extended Abstract.
21. Kinny, D. 1993. The distributed multi-agent reasoning system architecture and language specification. Technical report, Australian Artificial Intelligence Institute, Melbourne, Australia.
22. Krafta, R. 1999. Spatial self-organization and the production of the city. *Urbana* 24:49–62.
23. Leite, J. A., Alferes, J. J. and Pereira, L. M. 2002. $\mathcal{MINERVA}$—a dynamic logic programming agent architecture. In Meyer, J.-J. and Tambe, M., eds., *Intelligent Agents VIII – Proceedings of the Eighth International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001), August 1–3, 2001, Seattle, WA,* number 2333 in Lecture Notes in Artificial Intelligence, 141–157. Berlin: Springer-Verlag.
24. Lesser, V. R. 1998. Reflections on the nature of multi-agent coordination and its implications for an agent architecture. *Autonomous Agents and Multi-Agent Systems* 1(1):89–111.

25. Lòpez, F. and Luck, M. 2002. Towards a model of the dynamics of normative multi-agent systems. (In this volume).
26. Machado, R. and Bordini, R. H. 2002. Running AgentSpeak(L) agents on SIM_AGENT. In Meyer, J.-J. and Tambe, M., eds., *Intelligent Agents VIII – Proceedings of the Eighth International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001), August 1–3, 2001, Seattle, WA,* number 2333 in Lecture Notes in Artificial Intelligence, 158–174. Berlin: Springer-Verlag.
27. Moreira, Á. F. and Bordini, R. H. 2002. An operational semantics for a BDI agent-oriented programming language. In *Proceedings of the Workshop on Logics for Agent-Based Systems (LABS-02), held in conjunction with the Eighth International Conference on Principles of Knowledge Representation and Reasoning (KR2002), April 22–25, Toulouse, France,* 45–59.
28. Portugali, J. 2000. *Self-organization and the City.* Berlin: Springer-Verlag.
29. Rao, A. S. and Georgeff, M. P. 1995. BDI agents: From theory to practice. In Lesser, V. and Gasser, L., eds., *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95), 12–14 June, San Francisco, CA,* 312–319. Menlo Park, CA: AAAI Press / MIT Press.
30. Rao, A. S. and Georgeff, M. P. 1998. Decision procedures for BDI logics. *Journal of Logic and Computation* 8(3):293–343.
31. Rao, A. S. 1996. AgentSpeak(L): BDI agents speak out in a logical computable language. In Van de Velde, W. and Perram, J., eds., *Proceedings of the Seventh Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'96), 22–25 January, Eindhoven, The Netherlands,* number 1038 in Lecture Notes in Artificial Intelligence, 42–55. London: Springer-Verlag.
32. Russell, S. and Norvig, P. 1995. *Artificial Intelligence: A Modern Approach.* Prentice Hall Series on Artificial Intelligence. Upper Saddle River, NJ: Prentice Hall.
33. Shoham, Y. 1993. Agent-oriented programming. *Artificial Intelligence* 60:51–92.
34. Sloman, A. and Logan, B. 1999. Building cognitively rich agents using the SIM_AGENT toolkit. *Communications of the Association of Computing Machinery* 43(2):71–77.
35. Wagner, T., Garvey, A. and Lesser, V. 1998. Criteria-directed heuristic task scheduling. *International Journal of Approximate Processing, Special Issue on Scheduling* 19(1–2):91–118.
36. Wooldridge, M. J., Jennings, N. R. and Kinny, D. 2000. The Gaia methodology for agent-oriented analysis and design. *Autonomous Agents and Multi-Agent Systems* 3(3):285–312.
37. Wooldridge, M. 1999. Intelligent agents. In Weiß, G., ed., *Multiagent Systems—A Modern Approach to Distributed Artificial Intelligence.* Cambridge, MA: MIT Press. chapter 1, 27–77.

Organisation Modelling for the Dynamics of Complex Biological Processes

Tibor Bosse, Catholijn M. Jonker, and Jan Treur

Vrije Universiteit Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
{tbosse, jonker, treur}@cs.vu.nl
http://www.cs.vu.nl/~{tbosse, jonker, treur}

Abstract. This paper shows how an organisation modelling approach can be used to model the dynamics of biological organisation, in particular the circulatory system in biological organisms (mammals). This system consists of a number of components that are connected and grouped together. Dynamic properties at different levels of aggregation of this organisation model have been identified, and interlevel relationships between these dynamic properties at different aggregation levels were made explicit. Based on the executable properties simulation has been performed and properties have been checked for the produced simulation traces. Thus the logical relationships between properties at different aggregation levels have been verified. Moreover, relationships between roles within the organisation model and realisers of these roles have been defined. This case study shows that within biological and medical domains organisation modelling techniques can play a useful role in modelling complex systems at a high level of abstraction.

1 Introduction

In biological systems often many complex distributed interacting processes take place, that together result in some form of coherent joint action. Examples of such biological systems are mammals, insect colonies and bacteria. During evolution, Nature has developed several forms of organisational structure; typical examples are the organisation of a beehive, the coordinated processes of organs in mammals, and the well-organised regulated biochemistry of a living cell. Usually such biological systems are addressed by modelling the underlying physical/chemical processes by mathematical and system theoretical techniques, for example sets of differential equations; e.g., [26]. For some small unicellular organisms, a few isolated chemical pathways are understood in sufficient kinetic detail to obtain a description (by differential equations) of their import and primary processing of nutrients; e.g., in *Escherichia coli* [22], [24], or yeast [21]. However, even if all details would be available, at best this approach provides a description that is inherently low-level and complex. The adequacy of such mathematical techniques addressing the underlying physical/chemical level can be questioned. Such approaches do not exploit the apparent organisational structure that can be identified at a conceptual level within the biological systems addressed; the types of techniques often used are not tuned to

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 92–112, 2004.

modelling at such a conceptual level of the organisation of the distributed interacting processes.

In the area of organisation modelling, to handle complex distributed dynamics of the interaction between multiple agents in human society, often some type of organisational structure is exploited. The dynamics that emerge from multiple interacting agents within human society has been studied within Social Sciences in the area of Organisation Theory (e.g., [12], [13], [17], [19]) and within Artificial Intelligence in the area of Agent Systems (e.g., [2], [25]). To manage complex, decentralised dynamics in human society, organisational structure is a crucial element: organisation provides a structuring and co-ordination of the processes in such a manner that a process or agent involved can function in a more adequate manner. The dynamics shown by a given organisational structure is much more dependable than in an entirely unstructured situation. To exploit such organisational structures in a society particularly in modelling of these processes, within the agent systems area a number of conceptual modelling approaches have been developed, where a specific form of organisational structure is taken as a central concept. One of the recently developed organisational modelling approaches is the Agent/Group/Role (AGR) approach introduced in [3], extended with operational semantics in [4], and with a specification language for dynamic properties in [5].

Like in human societies, as discussed above, many biological systems take the form of complex organised distributed interacting processes. Therefore a natural research question addressed in this paper is whether organisational modelling techniques provide adequate means to model such biological systems at a conceptual-organisational level. If such an approach succeeds, it may be expected that it results in models of a much higher level than those addressing the biological processes at the level of their physiology or chemistry. A relating hypothesis is that such higher-level models can be simulated and analysed much more easily than the more complex mathematical models. These are the issues addressed in this paper. To explore these issues, in a rather arbitrary manner one specific available organisation modelling framework has been chosen and one specific organised biological phenomenon on which this organisation modelling framework was applied.

The chosen organisation modelling framework is the one described in [10], addressing both analysis and simulation of AGR-models, and supported by a software environment; a formal foundation can be found in [10]. This dynamic modelling environment allows to

- specify dynamic properties for the different elements and levels of aggregation within an AGR organisation model
- relate these dynamic properties to each other according to the organisational structure
- use dynamic properties in executable form as a declarative specification of a simulation model and perform simulation experiments
- automatically check dynamic properties for simulated or empirical traces

The goal of this paper is, in particular, to illustrate how this dynamic modelling framework for organisations, whilst being a conceptual approach, can also be used to model complex organised dynamics in biological systems involving several interacting processes.

The chosen case study for such a biological system, concerns the most primary dynamics of the circulatory system in biological organisms (mammals in particular).

This biological system shows sufficient complexity to be an interesting challenge. In the literature, many different kinds of cardiovascular (CV) models exist, typically based on modelling the physiology by differential equations. The first modern CV models were based on the *Windkessel* theory (the idea that arterial elasticity has a buffering effect on the pulsatile nature of blood flow), e.g. [16], [18], [20]. Another modern approach, that is influential in CV modelling today, makes use of hydrodynamic pulse-wave models [6], [10], [16], [18]. Furthermore, a distinction can be made between so-called transmission line models [27], segmental models [7], [15], [23], [28] and hybrid models. What all these approaches have in common is that they use rather complex models based on differential equations at the level of detailed physiology to describe the dynamics of this system.

In contrast, the current paper shows that the organisation modelling approach, although initially meant for purely social systems, provides adequate models in this type of application area as well. Realisers of roles within such an organisation models are active components of the biological system. As a result, this kind of biological organisations can also be considered in a way as (pseudo-)social systems, especially in the sense that the processes involved within these active components have to co-operate in a well-organised manner in order to produce the desired or required behavior for the overall system.

In Section 2 a brief introduction of the AGR organisation modelling approach can be found and illustrated for the context of the circulatory system. In Section 3 the dynamic properties at different levels of aggregation of this organisation model are identified. In Section 4 the relationships between these dynamic properties at different levels are presented. Section 5 describes how part of the dynamic properties can be used to enable a simulation of the circulatory system. In Section 6 the remaining properties are validated against the simulation of Section 5. Finally, Section 7 provides a description of how specific agents can be allocated to roles within the AGR approach.

2 The Organisation Structure of the Circulatory System

This section presents the organisation structure for the biological case study undertaken to investigate the usefulness of the AGR multi-agent organisation modelling approach to biological systems: the circulatory system in mammals. After a description of the functioning of the circulatory system, the AGR approach is briefly introduced. Next, the approach is applied to the circulatory system by identifying the organisational structure, expressed by AGR in terms of roles, groups, and interactions between these elements, and the agents realising these roles.

2.1 The Circulatory System

The circulatory system takes care for a number of capacities, such as providing nutrients and oxygen to the body and taking wastes (e.g., CO_2) out of the body; e.g., [18], [20]. The main property to focus on in this example is that the system provides oxygen for all parts of the body. The organisation of the circulatory system S is analysed as consisting of the following active components (or agents) that by showing their reactive and pro-active behavior all play their roles within the overall process:

- heart
- capillaries in lungs and other organs
- arteries
    - pulmonary artery channels (from the heart to the capillaries in the lungs)
    - aorta channels (from heart to the capillaries in the body)
- veins
    - pulmonary veins (from the capillaries in the lungs to the heart)
    - inferior and superior vena cava (from the capillaries in the body to the heart)

These active components work together due to some structure, as schematically depicted in Figure 1. Note that Figure 1 only describes the material structure of the circulatory system; the components depicted are physical components. Such pictures do not account for the role that the different physical components play in the organised process as a whole. For example the similarity in roles of the components in the systemic cycle (left hand side) and in the pulmonary cycle (right hand side) are not made precise. To clarify such functional and organisational aspects and similarities, the organisational structure will be described in the next subsections.

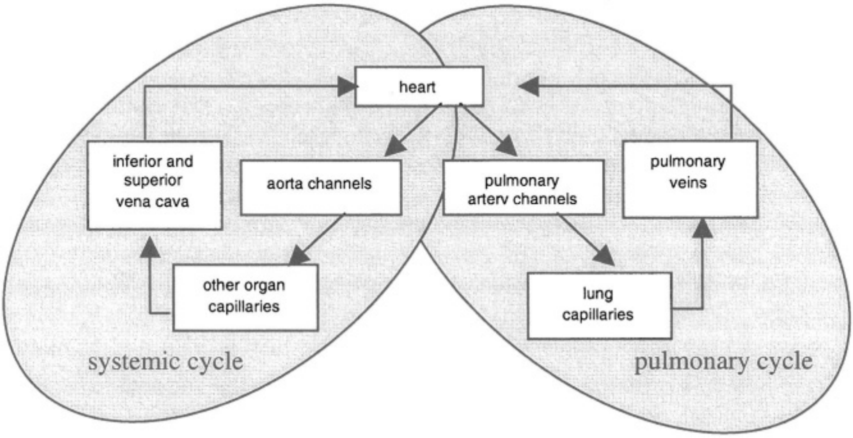


Fig. 1. Schema for the circulatory system

2.2 AGR Organisational Structures

To model an organisation, the Agent/Group/Role (AGR) approach, adopted from [3] is used. The *organisational structure* is the specification of a specific multi-agent organisation based on a definition of groups, roles and their relationships within the organisation:

- An organisation as a whole is composed of a number of *groups*.
- A group structure identifies the *roles* and (*intragroup*) *interaction between roles,* and *transfers* between roles needed for such interactions.
- In addition, *intergroup* role relations between roles of different groups specify the connectivity of groups within an organisation.

The modelling approach is further explained and illustrated by the application to the circulatory system in mammals.

2.3 Groups and Roles within the Circulatory System

The left-hand side and the right-hand side of the picture in Figure 1 are organised according to a similar structure:

- The *heart* initiates the flow,
- which is led by (aorta, resp. pulmonary artery) *arteries* or *channels* to
- *organs* (lung, resp. other organs) where exchange takes place,
- from where the flow is led by (pulmonary, resp. inferior and superior vena cava) *veins*
- back to the *heart.*

Here, in each of the two sides the heart plays two roles, one of a well, initiating the flow, and one of a drain, where the flow disappears (to re-appear in the other well).

The similarity of the two parts of the circulatory system enables to model their common structure in an abstract manner in the form of a more generic *group structure* G which has two instantiations within the circulatory system: one for the left hand side (called *systemic cycle,* used for oxygen supply, among others), and one for the right hand side (called *pulmonary cycle,* used for oxygen uptake, among others). Modelling the system from this perspective provides several advantages over the material perspective shown in Figure 1. For instance, the possibility to describe both main cycles by a single, generic group structure allows us to identify certain similarities between the two cycles. Moreover, such generic structures could enable comparative studies with systems in other organisms than mammals.

Generic Group Structure G

The generic group structure G (see Figure 2) consists of the following five *roles: well, supply guidance, exchange, drain guidance, drain.*

Transfers and Intragroup Role Interactions within G

The transfers underlying the interactions between roles are depicted in Figure 2. A short explanation of these interactions is as follows:

well – supply guidance role interaction

If the well comes up with a new flow, then this flow will be picked up by the supply guidance, and transported further.

supply guidance – exchange role interaction

If the supply guidance delivers a flow, then the exchange role will take out substances from this flow and will insert other substances in the flow.

exchange – drain guidance role interaction

The flow resulting from the exchange will be picked up by and transported by the drain guidance.

drain guidance – drain role interaction

If the drain guidance delivers a flow, then this is picked up by the drain (which lets it disappear).

Group Instances and Role Instances

Two instances of the generic group structure G are used: the *pulmonary cycle group instance* G_p and the *systemic cycle group instance* G_s. Based on the generic group structure G, for each of the group instances different role instances are defined. These role instances are denoted by using the group instance name as a prefix; i.e., the role instances *systemic cycle well, systemic cycle supply guidance, systemic cycle exchange, systemic cycle drain guidance, systemic cycle drain* within the systemic cycle group instance, and similar for the pulmonary group instance.

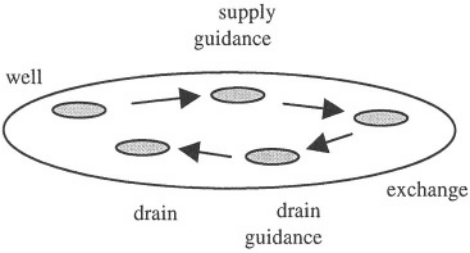


Fig. 2. Roles and transfers within the generic group structure G

Allocation of Agents to Role Instances

The relation between Figures 2 and 1 is that to each role instance depicted in Figure 2, a specific agent is allocated in Figure 1. This is the case for both the pulmonary cycle group instance and the systemic cycle group instance. In particular, for the systemic cycle group instance the allocation of agents to role instances is as follows:

heart	-	systemic cycle well
aorta channels	-	systemic cycle supply guidance
organ capillaries	-	systemic cycle exchange
inferior and superior vena cava	-	systemic cycle drain guidance
heart	-	systemic cycle drain

For the pulmonary cycle group instance the allocation of agents to role instances is:

heart	-	pulmonary cycle well
pulmonary channels	-	pulmonary cycle supply guidance
lung capillaries	-	pulmonary cycle exchange
pulmonary veins	-	pulmonary cycle drain guidance
heart	-	pulmonary cycle drain

The allocation of agents to role instances is discussed in more detail in Section 7.

2.4 Connectivity between Groups: Intergroup Role Interactions

The connectivity between the groups within the organisation structure is realised by two intergroup role interactions: from the drain role instance within one group to the well role instance in the other group, in both directions; see Figure 3.

In a generic sense such an intergroup role interaction can be explained by stating that the flow taken out by the drain role instance in one group instance is supplied within the other group instance by the well role instance. For the two group instances in the example these interactions are briefly explained as follows.

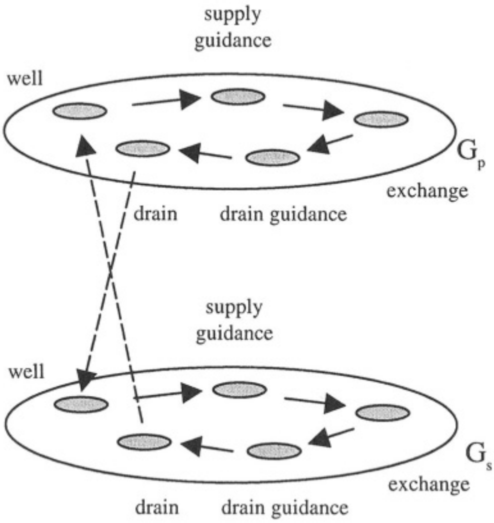


Fig. 3. Intergroup role interactions

- *pulmonary cycle drain – systemic cycle well role interaction*
  The oxygen-rich blood flow taken out by the pulmonary cycle drain role instance within the pulmonary cycle group instance is supplied to the systemic cycle well role instance within the systemic cycle group instance
- *systemic cycle drain – pulmonary cycle well role interaction*
  The oxygen-poor blood flow taken out by the systemic cycle drain role instance within the systemic cycle group instance is supplied to the pulmonary cycle well role instance within the pulmonary cycle group instance.

3 Dynamic Properties at Different Levels within the Organisation

To describe the functioning of the circulatory system S as an organisation, the following types of dynamic properties can be used (in the paper limited to properties related to oxygen supply which is a core function of the circulatory system):

- dynamic properties of the organisation as a whole
- dynamic properties for groups and intergroup role interactions
- properties of roles, transfer properties and intragroup role interactions within a group.

Moreover, usually some environmental assumptions are needed. The argument "s" when appearing in the name of a property refers to the instance of that property suitable for the systemic cycle group, similarly the argument "p" refers to the pulmonary cycle group.

3.1 Environment Assumptions

For the circulatory system S two reasonable environmental assumptions are:

EA1 **Oxygen availability**
At any point in time oxygen is present in the lungs

EA2(i) **Stimulus occurrence (with maximal interval i)**
For any point in time t there exists a time point with $t < t' \leq t + i$ such that at t' a stimulus occurs.

3.2 Dynamic Properties of the Organisation as a Whole

Global properties can be expressed for proper functioning of the flow through the cycles (taken at the well), and for resulting oxygen provision through the capillaries.

GP1(w) **Well successfulness (with maximal interval w)**
After an initiation time t0, for any point t there exists a time point t' with $t < t' \leq t + w$ such that at t' a fluid with ingredients I is generated by the well.

Here I is a specification of ingredients, for example by a list of them, possibly with indications of concentrations. Note that this global property depends on the organisation as a whole, not only on the group of the well. This property can be instantiated both for the well within the pulmonary cycle group $(GP1(p, w_p))$, and for the well within the systemic cycle group $(GP1(s, w_s))$.

GP2(d) **Oxygen delivery successfulness (with maximal interval d)**
After an initiation time t0, for any point t there exists a time point t' with $t < t' \leq t + d$ such that at t' by exchange oxygen is delivered to the organs.

3.3 Intergroup Role Interaction Properties

Intergroup role interaction properties relate roles in different groups. They typically express a dynamic relation between the input of one role in one group to the output of another role in another group. For the organisation of the circulatory system S consisting of two group instances as depicted in Figure 3 the following intergroup role interaction property has been specified. Again, this property can be instantiated both for the well within the pulmonary cycle group (IrRI(p, c_p, r_p)), and for the well within the systemic cycle group (IrRI(s, c_s, r_s)).

IrRI(c, r) Drain– well intergroup role interaction

At any point in time t0

if at some $t \leq t0$ the drain within some group instance G_i received a fluid volume V with ingredients I

and between t and t0 no stimulus occurred

and at t0 a stimulus occurs

then there exists a time point t 1 with $t0 + c \leq t1 \leq t0 + r$ such that at t 1 the well within the other group instance G_j generates a fluid volume V with ingredients I

3.4 Dynamic Properties of Groups

Within an overall organisation, each group's contribution can be formulated in the form of some group property. An example of such a group property is the following.

GR(u, v, u', v') Group successfulness

At any point in time t,

if at t the well generates a fluid volume V with ingredients I

then there exist time points $t' \leq t''$ with $t + u \leq t' \leq t + v$ and $t + u' \leq t'' \leq t + v'$ such that at t' ingredient A is added to the environment and ingredient B taken from the environment

and at t" the drain receives a fluid volume V with ingredients I - A + B

Here V is an amount of fluid and I is a specification of ingredients, as before. The notation I - A + B is used for the specification of the ingredients of I except A and augmented by B. The group specific property instances according to group instances are called GR(s, u_s, v_s, u'_s, v'_s) and GR(p, u_p, v_p, u'_p, v'_p). For the pulmonary group instance GR(p) the air is environment, A is carbonacid, and B is oxygen, for the systemic group instance GR(s) the environment is formed by the organs of the body, A is oxygen, and B is carbonacid. The difference in meaning of A and B for instantiations according to group instances is valid in other properties as well.

The dynamic properties of the different groups and of their interactions modelled by intergroup role interactions, contribute to the overall properties of S. As discussed in [5], some dynamic group properties have a specific form in that they relate one role in the group to another role in the group. The two types of such properties that are relevant (transfer properties and intragroup role interaction properties) are discussed in the following section.

3.5 Transfer and Intragroup Role Interaction Properties

Intragroup role interaction properties characterise how roles (have to) interact. They typically relate the output of one role to the output of another role. This is slightly more abstract than role behavior and transfer properties.

IaRI(a1, b1) Well implies supply guidance
At any point in time t
if the well generates a fluid volume V with ingredients I
then there exists a time point t' with $t + a1 \leq t' \leq t + b1$ such that at t'
the supply guidance generates a fluid volume V with ingredients I

IaRI2(a2, b2) Supply guidance implies exchange
At any point in time t
if the supply guidance generates a fluid volume V with ingredients I
then there exists a time point t' with $t + a2 \leq t' \leq t + b2$ such that at t'
ingredient A is added to the object and ingredient B taken from the object
and the exchange generates a fluid volume V with ingredients I - A + B

IaRI3(a3, b3) Exchange implies drain guidance
At any point in time t
if the exchange generates a fluid volume V with ingredients J
then there exists a time point t' with $t + a3 \leq t' \leq t + b3$ such that at t'
the drain guidance generates a fluid volume V with ingredients J

Transfer properties express that the different roles are connected in an appropriate manner to enable proper interaction. For each of the four arrows in Figure 3 a transfer property expresses that the proper connection exists between the output of one role and the input of the other role. In a general form delays can be taken into account for the transfers. However, for this example, these delays for transfers are assumed to be 0 (input state property is assumed identical to previous output state property), i.e., all gi’s and hi’s are 0.

TR1(g1, h1) Well connects to supply guidance
At any point in time t
if the well generates a fluid volume V with ingredients I
then there exists a time point t' with $t + g1 \leq t' \leq t + h1$ such that at t'
the supply guidance receives a fluid volume V with ingredients I

This property is not fulfilled, for example, if the well opening is not connected to the supply guidance, so that the generated fluid volume streams away in the environment without reaching the supply guidance.

TR2(g2, h2) Supply guidance connects to exchange
At any point in time t
if the supply guidance generates a fluid volume V with ingredients I
then there exists a time point t' with $t + g2 \leq t' \leq t + h2$ such that at t'
the exchange receives a fluid volume V with ingredients I

TR3(g3, h3) **Exchange connects to drain guidance**
At any point in time t
if the exchange generates a fluid volume V with ingredients I
then there exists a time point t' with $t + g3 \leq t' \leq t + h3$ such that at t' the drain guidance receives a fluid volume V with ingredients I

TR4(g4, h4) **Drain guidance connects to drain**
At any point in time t
if the drain guidance generates a fluid volume V with ingredients I
then there exists a time point t' with $t + g4 \leq t' \leq t + h4$ such that at t' the drain receives a fluid volume V with ingredients I

3.6 Role Behavior Properties

Role behavior properties abstract from the specific agent allocated to a role, but characterise which behavior an agent fulfilling this role needs to have. Such properties typically relate the input of a role to the output of the same role.

supply guidance behavior

The arteries contribute in transportation. This means that that if their input receives blood, then their output generates blood with the same ingredients.

RB1(e1, f1) **Supply guidance effectiveness**
At any point in time t
if the supply guidance receives a fluid volume V with ingredients I
then there exists a time point t' with $t + e1 \leq t' \leq t + f1$ such that at t' it generates a fluid volume V with ingredients I

exchange behavior

RB2(e2, f2) **Exchange effectiveness**
At any point in time t
if the exchange receives a fluid volume V with ingredients I
then there exists a time point t' with $t + e2 \leq t' \leq t + f2$ such that at t' ingredient A is added to the object (environment, i.e., lung or organ)
and ingredient B is taken from the object
and it generates a fluid volume V with ingredients I - A + B

drain guidance behavior

RB3(e3, f3) **Drain guidance effectiveness**
At any point in time t
if the drain guidance receives a fluid volume V with ingredients I
then there exists a time point t' with $t + e3 \leq t' \leq t + f3$ such that at t' it generates a fluid volume V with ingredients I

4 Relationships between Dynamic Properties at Different Levels

The idea is that dynamics of the whole organised (multi-agent) system is generated by lower level properties, in particular by the group properties and intergroup interaction properties. In turn, group dynamics is generated by role behavior and transfer within a group. This is elaborated in more detail by identifying logical relationships between these dynamic properties.

4.1 Overall Properties: Oxygen Delivery Successfulness

The global property GP2 (oxygen delivery Successfulness) depends on the systemic cycle instance of global property GP1 (well Successfulness), assuming proper group functioning of the same group instance. To be more precise, the following relationship holds:

$$\mathrm{GP1}(s, w) \;\&\; \mathrm{GR}(s, u_s, v_s, u'_s, v'_s) \Rightarrow \mathrm{GP2}(d)$$

$$\textbf{with } d = w + v_s.$$

So property GP2(d) is implied by two other properties, i.e., GP1(s, w) and $\mathrm{GR}(s, u_s, v_s, u'_s, v'_s)$. This implication are depicted in Figure 4. A sketch of a proof of this implication is as follows. Suppose GP1(s, w) holds. Then, after an initiation time t0, for any point t there exists a time point t' with $t < t' \leq t + w$ such that at t' a fluid with ingredients I is generated by the well of the systemic cycle. And if $\mathrm{GR}(s, u_s, v_s, u'_s, v'_s)$ holds as well, this means that the systemic cycle works correctly. Thus, from the fluid generated by the well, oxygen is finally taken and delivered to the organs. It can be concluded that after an initiation time t0, for any point t there exists a time point t' with $t < t' \leq t + d$ such that at t' by exchange oxygen is delivered to the organs, which is exactly what GP2(d) states. Furthermore, it is known that w is the maximum time interval for fluid generation by the well, and v_s is the maximum time interval for oxygen supply by the systemic cycle. Hence, it follows logically that $d = w + v_s$. The relationships that GP1(s, w) and $\mathrm{GR}(s, u_s, v_s, u'_s, v'_s)$ have with other properties are depicted in Figures 5 and 6.

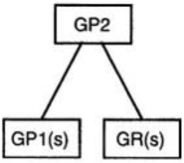


Fig. 4. Oxygen delivery successfulness related to global property GP1(s) and a group property.

4.2 Overall Properties: Well Successfulness

Well successfulness depends on proper functioning of the whole cycle; it needs as input that a fluid volume is received. If the whole cycle functions well, the group properties, intergroup role interaction properties, and environmental assumption EA2 guarantee that this well functioning is maintained. However, the process needs a starting point. This starting point is assumed for the well within both groups at time point t = 0 in the following form:

Init(w_{init}) **Well initialisation**

There exists a time point t with $0 \leq t \leq w_{init}$ such that at t
the well in the pulmonary group instance generates a fluid volume V with any ingredients I
and the well in the systemic group instance generates a fluid volume V' with any ingredients I'

Using these properties the following relationships can be established (see also Fig. 5).

$$\text{Init}(w_{init}) \ \&\ \text{GR}(s, u_s, v_s, u'_s, v'_s) \ \&\ \text{GR}(p, u_p, v_p, u'_p, v'_p) \ \&\ \text{IrRI}(s, c_s, r_s) \ \&\ \text{IrRI}(p, c_p, r_p) \ \&\ \text{EA2}(i) \Rightarrow \text{GP1}(s, w_s)$$

with $w_s = \max(w_{init}, \max(i, v'_p) + r_s)$.

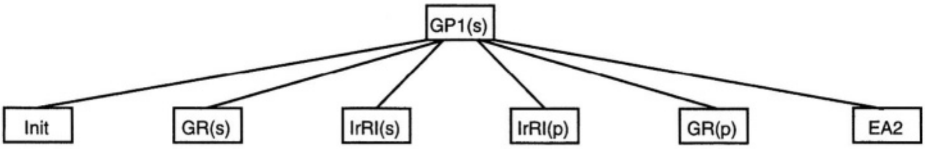


Fig. 5. Global property GP1(s) related to other properties

4.3 Group Properties

A group property is related in an integrative manner to a combination of intragroup role interaction properties.

$$\text{IaRI1}(s, a1_s, b1_s) \ \&\ \text{IaRI2}(s, a2_s, b2_s) \ \&\ \text{IaRI2}(s, a3_s, b3_s) \Rightarrow \text{GR}(s, u_s, v_s, u'_s, v'_s)$$

with $u_s = a1_s + a2_s$, $v_s = b1_s + b2_s$, $u'_s = a1_s + a2_s + a3_s$, $v'_s = b1_s + b2_s + b3_s$.

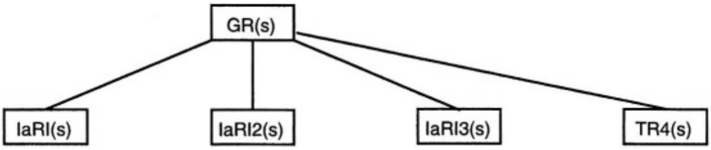


Fig. 6. Group property related to intragroup interaction properties

Intragroup role interaction properties relate to role behavior properties and transfer properties in the following manner.

$$\mathrm{TR1(s)\ \&\ RB1_s(s,\ e1_s,\ f1_s)\ \Rightarrow\ IaRI1(s,\ e1_s,\ f1_s)}$$
$$\mathrm{TR2(s)\ \&\ RB2(s,\ e2_s,\ f2_s)\ \Rightarrow\ IaRI2(s,\ e2_s,\ f2_s)}$$
$$\mathrm{TR3(s)\ \&\ RB3(s,\ e3_s,\ f3_s)\ \Rightarrow\ IaRI3(s,\ e3_s,\ f3_s)}$$

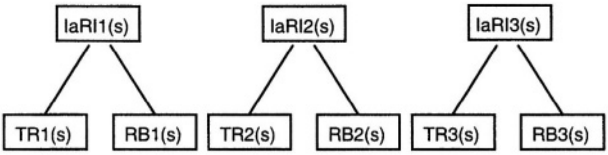


Fig. 7. Intragroup interaction properties related to role behavior and transfer properties

4.4 Overview

In Figure 8 an overview can be found for all dynamic properties relating to $\mathrm{GP1_s}$.

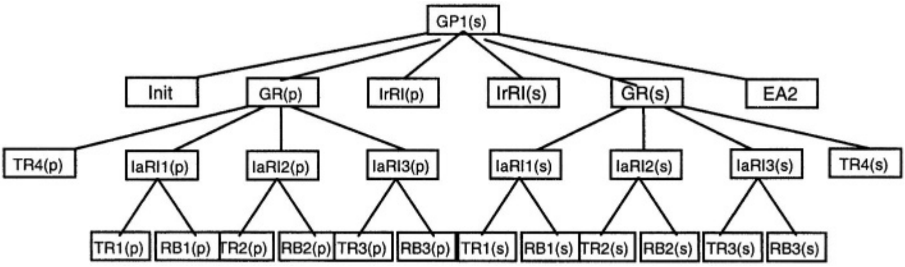


Fig. 8. Overview of the interlevel relationships for global property GP1(s)

5 Simulation

A software environment has been created to enable the simulation of executable organisation models specified at a high conceptual level [10]. The input of this simulation environment is a set of dynamic properties in a specific, executable format. In [9] the language TTL was introduced as an expressive language for the purpose of specification and checking of dynamic properties. For the purpose of simulation, to obtain computational efficiency the format used for dynamic properties is more restricted than the TTL format used to specify various types of dynamic properties: they are in so-called *leads to* format; cf. [10]. This is a real time-valued variant of Executable Temporal Logic [1]. Roughly spoken, in *leads to* format the following can be expressed:

ifa state property α *holdsfor a time interval with duration g,*
then after some delay (between e andf) another state property β *will holdfor a time interval h*

This specific temporal relationship *leads to* is applicable forward as well as backward in time. Hence, if α and β are state properties, and α leads to β, this also means that if β holds for a time interval of length h, then α held during some time interval with length g, of which the starting point was between e and f before the starting point of the second interval. A formal definition of this *leads to* relation is as follows. Here $\mathrm{state}(\mathcal{T}, t)$ denotes the state at time t in trace $\mathcal{T}$, and $S \models \alpha$ that in a state S state property α holds. Moreover, Traces denotes the set of all possible traces.

Definition

(a) Let α, $\beta \in \mathrm{SPROP(AllOnt)}$. The state property α *follows* state property β, denoted by $\alpha \twoheadrightarrow_{e,\,f,\,g,\,h} \beta$, with time delay interval [e, f] and duration parameters g and h if

$\forall \mathcal{T} \in \mathrm{Traces}\ \ \forall t1:$

$[\forall t \in [t1 - g, t1) : \mathrm{state}(\mathcal{T}, t) \models \alpha \ \Rightarrow\ \exists d \in [e, f]\ \forall t \in [t1 + d, t1 + d + h) : \mathrm{state}(\mathcal{T}, t) \models \beta\,]$

(b) Conversely, the state property β *originates from* state property α, denoted by $\alpha \bullet\!\!-_{e,\,f,\,g,\,h} \beta$, with time delay in [e, f] and duration parameters g and h if

$\forall \mathcal{T} \in \mathrm{Traces}\ \forall t2:$

$[\forall t \in [t2, t2 + h) : \mathrm{state}(\mathcal{T}, t) \models \beta \Rightarrow \exists d \in [e, f]\ \forall t \in [t2 - d - g, t2 - d)\ \ \mathrm{state}(\mathcal{T}, t) \models \alpha]$

(c) If both $\alpha \twoheadrightarrow_{e,f,g,h} \beta$, and $\alpha \bullet\!\!-_{e,f,g,h} \beta$ hold, then α *leads to* β this is denoted by: $\alpha \bullet\!\!\twoheadrightarrow_{e,f,g,h} \beta$.

Making use of these *leads to* properties, the software environment generates simulation traces (actually the *follows* relations are used in the simulation software; if in a specification there is only one way to reach each β, then this automatically results in *leads to* relations holding). A trace is developed by starting at time t = 0 and for each time point up to which the trace already has been constructed, checking which antecedents of executable properties hold in the already constructed trace. For these executable properties, add the consequent to the trace, i.e., extend the trace in time in such a manner that the consequent holds.

The relation between the specification and the constructed trace is that the trace is a model (in the logical sense) of the theory defined by the specification, i.e., all executable dynamic *leads to* properties of the specification hold in the trace.

To be able to simulate the behavior of the circulatory system, all leaves of the tree in Figure 8 have been expressed in *leads to* format. That is, all intergroup role interaction properties, role behavior properties, transfer properties, and the special starting point property Init. The values chosen for the timing parameters are shown in Table 1.

The resulting trace is shown in Figure 9. Time is on the horizontal axis, the properties are on the vertical axis. A dark box on top of the line indicates that the property is true during that time period, and a lighter box below the line indicates that the property is false during that time period. The line labeled *stimulus_occurs,* for example, depicts the property that a heart stimulus occurs. This property is true from time point 0 to 5, from 80 to 85, from 160 to 165, and so on. Notice that this is exactly the intended dynamics according to environmental assumption EA2. Also notice that for the maximum interval s within EA2, the value 80 has been chosen within this example. Furthermore, Figure 9 shows that after a stimulus has occurred, the wells of both groups generate fluid, which is immediately received by the supply

Table 1. Time parameters for *leads to* properties

Property	Minimal delay (e)	Maximal delay (f)	Duration antecedent (g)	Duration consequent (h)
RB1(p)	3	5	0	0
RB1(s)	10	20	0	0
RB2(p)	5	10	0	0
RB2(s)	5	10	0	0
RB3(p)	3	5	0	0
RB3 (s)	10	20	0	0
IrRI(p)	5	10	1	10
IrRI(s)	5	10	1	10

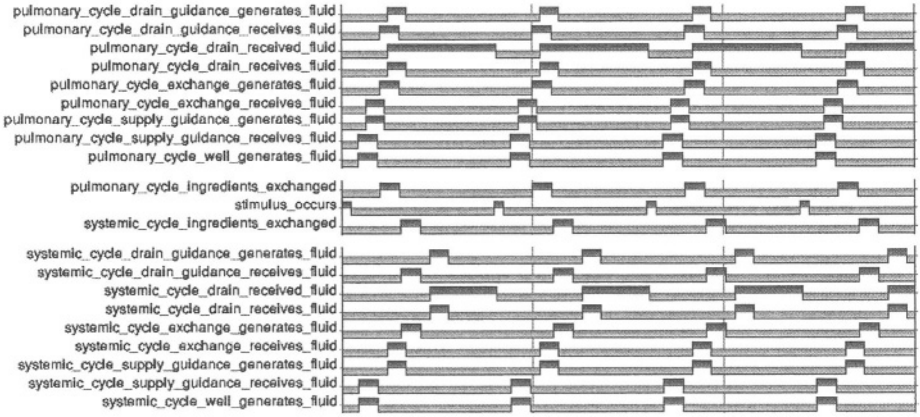


Fig. 9. Results of the simulation of executable properties of the circulatory system

guidances (since the delays for transfers were assumed to be 0). After that, in both groups the fluid continues to the exchange. Since the systemic cycle is longer than the pulmonary cycle (the aorta channels are longer than the pulmonary artery channels), it takes more time for the supply guidance in the systemic group to generate fluid. Next, some moments after the exchange has received a fluid, it can be seen that the ingredients are actually exchanged. After that, fluid goes from the exchange to the drain guidance and finally to the drain.

6 Checking Properties

Logical relationships between properties, as depicted in the tree of Figure 8, can be very useful in the analysis of dynamic properties of an organisation (like the circulatory system in this particular case); also see [8]. For example, if for a given trace of the system the global property GP1(s) is not satisfied, then by a refutation process it can be concluded that either one of the group properties, or one of the intergroup role interaction properties, or the property Init does not hold. If, after checking these properties, it turns out that GR(p) does not hold, then either one of the intragroup role interaction properties or TR4(p) does not hold. By this refutation analysis it follows that if GP1(s) does not hold for a given trace, then, via the intermediate properties, the cause of this malfunctioning can be found in the set of leaves of the tree of Figure 8.

In order to determine which one of the properties encountered in this refutation process actually is refuted, some mechanism is needed to check if a certain property holds for a given trace. To this end, the simulation software described above automatically produces log files containing the traces. In addition, software has been developed that is able to read in these log files together with a set of dynamic properties (in *leads to* format), and to perform the checking process. This is done in two directions. On the one hand, each atom occurring in the trace is 'explained', i.e., the software verifies if there was a reason for its presence, according to the dynamic properties. On the other hand, for each atom a check is performed whether all atoms it implies according to the dynamic properties are actually there. As a result, the software determines not only whether the properties hold for the trace or not, but in case of failure, it also pinpoints which parts of the trace violate the properties. If a property does not hold completely, this is marked by the program. Yellow marks indicate unexpected events, occurring when certain atoms cannot be explained. Red marks indicate events that have not happened, whilst they should have happened. Checks of this kind have actually been performed for all of the higher level properties of Figure 8, i.e., for all nodes of the tree that are no leaves. They all turned out to hold for the trace of Figure 9, which validates the tree.

In addition, recently other software has been developed (and is still being improved) that is able to check traces against properties in the *TTL* format instead of the *leads to* format. Since TTL, as mentioned in Section 5, has a considerably higher expressiveness, this new software enables to check much more complex properties. For instance, for the present case study, the property *"the higher the number of stimuli, the more oxygen is delivered in the lungs"* has been checked successfully. Checks of this kind are normally performed in less than a second. Future work involves exploring the limits to the amount of complexity that the software can handle.

7 Realisation of the Organisation by Allocation of Agents

An organisation model such as the one presented in this paper provides an abstract model for the manner in which multiple interacting processes or agents generate dynamics. The specific agents are not part of such an organisation model. Instead the notion of role provides an abstract entity or placeholder for where specific agents come in. In the example domain addressed here these agents are active biological components such as the heart, lungs, and other organs. An important advantage of this abstraction is that the dynamics can be modeled independent of the specific choices of agents. The organisation model can be (re)used for any allocation of agents to roles for which:

- for each role, the allocated agent's behavior satisfies the dynamic role properties,
- for each intergroup role interaction, one agent is allocated to both roles and its behavior satisfies the intergroup role interaction properties, and
- the communication between agents satisfies the respective transfer properties.

Expressed differently, for a given allocation of agents to roles the following logical relationships between dynamic properties hold:

agent – role
from dynamic agent properties to dynamic role properties:

> agent A is allocated to role r &
> dynamic properties of agent A ⇒
> dynamic properties of role r

As an example for the case of the circulatory system, one can think of the aorta channels as agent A and of the systemic cycle supply guidance as role r (also see the allocation schema at the end of Section 2.3).

agent – intergroup role interaction
from dynamic agent properties to dynamic intergroup role interaction properties:

> agent A is allocated to roles r1 and r2 in different groups &
> dynamic properties of agent A ⇒
> dynamic properties of intergroup role interaction between r1 and r2

As an example, one can think of the heart as agent A and of the systemic cycle well and the pulmonary cycle drain as role r1 and r2, respectively.

agent communication – role transfer
from dynamic agent communication properties to dynamic transfer properties:

> agent A is allocated to role r1 and agent B to role r2 in one group &
> dynamic properties of communication from A to B ⇒
> dynamic properties of transfer from r1 to r2

As an example, one can think of the aorta channels as agent A, of the systemic cycle supply guidance as role r1, of the organ capillaries as agent B and of the systemic cycle exchange as role r2.

8 Discussion

The aim of this paper was to investigate whether modelling techniques from the area of organisation modelling (already shown to be successful for human organisations in, e.g., [8], [11]) provide adequate means to model at a high level of abstraction the dynamics of biological systems in which multiple distributed interacting processes play a role. As a case study the circulatory system in biological organisms (mammals) was explored using a chosen organisation modelling framework.

In the literature, many different kinds of cardiovascular models exist, typically based on modelling the physiology by differential equations. In contrast to these mathematical models of the circulatory system our paper shows how an organisation modelling approach such as the chosen one (other organisation modelling approaches may well be as applicable as the chosen one) can be used to model the dynamics of biological organisation for the case of the circulatory system at a high conceptual level. This system consists of a number of components that are connected and grouped together in such a manner that everything functions in a coherent manner. It was shown how active components within the circulatory system can be considered realisers of the roles within the organisation model. Dynamic properties at different levels of aggregation of this organisation model have been identified, and logical interlevel relationships between these dynamic properties at different aggregation levels were made explicit. Based on the executable properties, simulation has been performed and properties have been (automatically) checked for the produced simulation traces. Thus the logical interlevel relationships between properties have been verified. The variant of executable temporal logic (extending the approach described in [1]) used for simulation has as an advantage that it is guaranteed that a generated trace satisfies the specified executable dynamic properties. Since these dynamic properties stand in logical relationships to other (more complex, not necessarily executable) dynamic properties, this form of simulation facilitates logical analysis of the dynamics at different levels of aggregation.

In summary, it turned out that, at least for the chosen domain, the chosen organisation modelling approach provides adequate means for high-level modelling of the complexity of the dynamics of biological organisms. For example, a strong contrast in abstraction and manageability of the model was found with modelling techniques based on differential equations that provide less transparent, low-level models. This outcome was confirmed by a case study in another biological domain in which the organisation of intracellular processes was modelled.

References

1. Barringer, H., Fisher, M., Gabbay, D., Owens, R., and Reynolds, M., (1996). The Imperative Future: Principles of Executable Temporal Logic, Research Studies Press Ltd. and John Wiley & Sons.
2. Ferber, J., (1999). *Multiagent Systems.* Addison Wesley.
3. Ferber, J. and Gutknecht, O. (1998). A meta-model for the analysis and design of organisations in multi-agent systems. In: *Proceedings of the Third International Conference on Multi-Agent Systems* (ICMAS'98), IEEE Computer Society Press, pp. 128–135.

4. Ferber, J., and Gutknecht, O. (1999). Operational Semantics of a role-based agent architecture. *Proceedings of the 6th Int. Workshop on Agent Theories, Architectures and Languages (ATAL'1999).* In: Jennings, N.R. & Lesperance, Y. (eds.) *Intelligent Agents VI,* Lecture Notes in AI, vol. 1757, Springer Verlag, 2000, pp. 205–217.
5. Ferber, J., Gutknecht, O., Jonker, CM., Müller, J.P., and Treur, J., (2001). Organization Models and Behavioural Requirements Specification for Multi-Agent Systems. In: Y. Demazeau, F. Garijo (eds.), *Multi-Agent System Organisations. Proceedings of the 10th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'01,* 2001. Lecture Notes in AI, Springer Verlag. To appear, 2002.
6. Fung, Y.C. (1984). Biodynamics: Circulation, New York, Springer-Verlag.
7. Greenway, C.V. (1982). Mechanisms and quantitative assessment of drug effects on cardiac output with a new model of the circulation. Pharm Rev 33(4):213.
8. Jonker, CM., Letia, I.A., and Treur, J., (2002). Diagnosis of the Dynamics within an Organisation by Trace Checking of Behavioural Requirements. In: Wooldridge, M., Weiss, G., and Ciancarini, P. (eds.), *Proc. of the 2nd International Workshop on Agent-Oriented Software Engineering, AOSE'01.* Lecture Notes in Computer Science, vol. 2222. Springer Verlag, 2002, pp. 17–32.
9. Jonker, C.M. and Treur, J. (1998). Compositional Verification of Multi-Agent Systems: a Formal Analysis of Pro-activeness and Reactiveness. In: W.P. de Roever, H. Langmaack, A. Pnueli (eds.), Proceedings of the International Workshop on Compositionality, COMPOS'97. Lecture Notes in Computer Science, vol. 1536, Springer Verlag, 1998, pp. 350–380. Extended version in: *International Journal of Cooperative Information Systems,* vol. 11, 2002, pp. 51–92.
10. Jonker, C.M., and Treur, J., Relating Structure and Dynamics in an Organisation Model. In: J.S. Sichman, F. Bousquet, and P. Davidson (eds.), *Multi-Agent-Based Simulation II, Proc. of the Third International Workshop on Multi-Agent Based Simulation, MABS'02,* Lecture Notes in AI, vol. 2581, Springer Verlag, pp. 50–69.
11. Jonker, C.M., Treur, J., and Wijngaards, W.C.A. (2002). Temporal Languages for Simulation and Analysis of the Dynamics Within an Organisation. In: B. Dunin-Keplicz and E. Nawarecki (eds.), *From Theory to Practice in Multi-Agent Systems, Proc. of the Second International Workshop of Central and Eastern Europe on Multi-Agent Systems, CEEMAS'01,* 2001. Lecture Notes in AI, vol. 2296, Springer Verlag, 2002, pp. 151–160.
12. Kreitner, R., and Kunicki, A. (2001). Organisational Behavior, McGraw – Hill.Li, J.K.-J. (1987). Arterial System Dynamics, NewYork, New York University Press.
13. Lomi, A., and Larsen, E.R. (2001). Dynamics of Organizations: Computational Modeling and Organization Theories, AAAI Press, Menlo Park.
14. Manna, Z., and Pnueli, A. (1995). *Temporal Verification of Reactive Systems: Safety.* Springer Verlag.
15. Martin, J.F., Schneider, A.M., Mandel, J.E., et al. (1986). A new cardiovascular model for real-time applications. Trans Soc Comp Sim 3(1):31.
16. Milnor, W.R. (1989). Hemodynamics 2nd ed, Baltimore, Williams & Wilkins.
17. Mintzberg, H. (1979). The Structuring of Organisations, Prentice Hall, Englewood Cliffs, N.J.
18. Noordergraaf, A. (1978). Circulatory System Dynamics. Academic Press, New York.
19. Prietula, M., Gasser, L., Carley, K. (1997). Simulating Organizations. MIT Press.
20. Rideout, V.C. (1991). Mathematical and Computer Modelling of Physiological Systems. Prentice Hall, Englewood Cliffs.
21. Rizzi, M., Baltes, M., Theobald, U. & Reuss, M. (1997). *In vivo* analysis of metabolic dynamics in *Saccharomyces cerevisiae:* II. Mathematical model. *Biotechnol. Bioeng.* 55, 592–608.
22. Rohwer, J.M., Meadow, N.D., Roseman, S., Westerhoff, H.V., Postma, P.W. (2000). Understanding glucose transport by the bacterial phosphoenolpyruvate:glycose phosphotransferase system on the basis of kinetic measurements in vitro. J. Biol. Chem., 275(45):34909–21.

23. Slate, J.B., Sheppard, L.C. (1982). Automatic control of blood pressure by drug infusion. IEE Proc., Pt. A 129(9):639.
24. Wang, J., Gilles, E.D., Lengeler, J.W., and Jahreis, K. (2001). Modeling of inducer exclusion and catabolite repression based on a PTS-dependent sucrose and non-PTS-dependent glycerol transport systems in Escherichia Coli K-12 and its experimental verification. *J. Biotechnol.* 2001, Dec 28; 92(2): 133–58.
25. Weiss, G. (ed.) (1999). *Multiagent Systems.* MIT Press
26. Westerhoff, H.V. (2001) The silicon cell, not dead but live! Metab. Eng. 2001; 3(3):207–10.
27. Westerhof, N., Noordergraaf, A. (1969). Reduced models of the systemic arteries. Proc 8th Int Conf Med Eng, Chicago.
28. Yu, C., Roy, R.J., Kaufman, H. (1990). A circulatory model for combined nitroprusside-dopamine therapy in acute heart failure. Med Prog Tech 16:77.

Communication without Agents? From Agent-Oriented to Communication-Oriented Modeling

Thomas Malsch[1] and Christoph Schlieder[2]

[1] Technical University Hamburg-Harburg,
Department of Technology Assessment
Schwarzenbergstr. 95, 21071 Hamburg, Germany
malsch@tuhh.de
[2] Bamberg University
Laboratory for Semantic Information Technology
christoph.schlieder@wiai.uni-bamberg.de

Abstract. From News to Chat, electronic discussion groups are widely acknowledged as a popular medium of communication. Unlike electronic mail which is rather easy to handle since it operates on a one-to-one bases, to keep up with forum discussions is extremely demanding. Participation in forums requires a constant effort of selection and attention from the user which goes beyond the limits of cognitive capacities. In this paper, we suggest to cope with this problem by introducing communication-oriented modeling (COM) as an alternative to agent-oriented modeling (AOM). Our approach to COM is based on theoretical foundations inspired by socionics and sociology.

1 Introduction

In this paper, we introduce a new approach into socionics and multi-agent systems research and design: *communication-oriented modeling* (COM). This methodological framework complements and reinforces *agent-based modeling* (AOM). In large-scale communication processes, especially those running on the Internet like discussion groups or chats, interaction between participants is often not organized along the lines of agent-to-agent relations. Rather, we find patterns of communication organized along the lines of message-to-message relations. Specifically, this can be observed in Internet-based public debates shaped by interrelated messages where a widely shared argumentation or a common view on a topic of general interest is gradually established [16].

In Internet discussion groups messages usually are not sent to a specific receiver but “To Whom It May Concern”. Messages are published to attract general attention and to enhance their social visibility by referring to other messages. Visibility in a general social sense, i.e. accessibility of a message and its potential of generating sequel messages, is a prerequisite of analyzing communication processes in real and artificial societies. Whenever a message is published for an audience rather than sent to a receiver, and whenever communication is dominated by messages referring to

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 113–133, 2004.

other messages rather than agents influencing other agents' beliefs, intentions, and actions, it is communications rather than agents that should be considered as the foundational units of analysis and modeling.

The remainder of this paper is organized as follows: Section 2 gives an overview of practical problems arising in Internet discussion groups as seen from the perspective of participants and moderators. This is to illustrate the need to reinforce multi-agent platforms with methods and tools based on a new communication-oriented approach that does not depend on speech act theory along the lines of Austin, Searle and Habermas. In section 3 we introduce *communication-oriented modeling* (COM) as a methodological concept based on a social theory of evolving networks of communication which assumes that society consists of communication events rather than human beings. Section 4 elaborates the technicalities of COM in well known formalisms for logical and graphical representation. Section 5 gives an outlook on future research together with some hints on how to apply COM to multi-agent systems. Finally, to highlight the originality of our approach, section 6 gives an overview of related work in DAI, sociology and socionics.

2 Speech Acts and Agent-Oriented Modeling

A prominent example for a communication process on the Internet – attractive to computer novices and experts alike – are the Usenet discussion groups which started in 1979, that is, more than a decade before the WWW. In December 2001, Google Inc. made its Usenet archive publicly available, thereby opening an incredibly rich source for studies in the history of computing and the sociology of communication. With more than 700.000.000 messages posted over a period of 20 years, the archive contains the best-documented large-scale communication process of the digital age. This makes it an ideal domain for illustrating some obvious limitations of *agent-oriented modeling* (AOM) of communication processes at a very large scale.

Agent-platforms such as FIPA-OS[1] [8] or JADE[2] are designed to enable communication between software entities (agents) showing goal-directed behavior rooted in a complex motivational system (BDI architectures). In the past few years, considerable effort has been made to develop agent communication languages [24, 28] in order to provide multi-agent systems with more transparency and coordinative power. The advantage of using agent-platforms for modeling communicative processes lies in the technical framework for agent communication which supports the purpose of communication analysis as well as that of simulation. An obvious way to model a Usenet discussion group within an agent platform consists in representing authors who are posting messages by agents of the platform. For modeling the messages, the unchallenged paradigm of communication in DAI is adopted: speech acts. Accordingly, along the lines of KQML standards [7], agent communication is conceptualized as an illocutionary act of a speaker (sender) who sends a message aiming at influencing the

[1] FIPA-compliant open source platform distributed by Emorphia Inc. (fipa-os.sourceforge.net).

[2] FIPA-compliant open source platform distributed by TILAB Inc. (www.telecomitalia.it).

addressee's (receiver) intentions and actions. A similar sender-receiver pattern dominates computer communication in which the exchange of messages is regulated by protocols describing the precise conditions of starting the communication, acknowledging receipt, and so on .[3] HTTP, the basic protocol of the WWW, for instance, adheres to the sender-receiver pattern as it regulates the flow of messages between a WWW client and a web server identified by an address, the URL [6].

In AOM, the minimal structure of a speech act can be described as being composed out of three components:

Table 1. The speech act in agent-oriented modeling

Agent1	Sender	Persistent
Agent2	Addressee	Persistent
Speech Act	Act(Proposition)	Transient

At a more complex level, communication processes are specified by interaction protocols which are composed out of sequences of several speech acts. Thus, interaction protocols account for the fact that the addressee of a speech act is an autonomous agent too. It is realistically assumed that both, sender and receiver, are taking turns (quite in line with turn taking in conversation analysis). Platforms such as FIPA-OS are equipped with specific interaction protocols for different types of communication processes relevant in distributed problem solving (e.g. contract net protocol).

Although AOM is very successful with regard to distributed and cooperative problem solving, its speech-act based conceptualization of communication follows the sender-receiver pattern and, as a consequence, it has to struggle with a number of conceptual deficiencies and shortcomings when applied to large-scale communication processes beyond the scope of small-group interaction. In the following we will highlight three problems of the message sending paradigm underlying AOM.

1. *Focus on agent-agent relations.* In AOM, it is the agent who is considered to be the driving force of communication. The primary task of modeling consists in representing which agent authors a message (sender) and which agent interprets that message (receiver). Related design question are: What is an agent's intention and how is it encoded in a message? However, in large-scale communication processes such as Usenet discussion groups, the intentional stance needs to be reinforced, if not substituted, with what we may call the referential or receptional stance: How is a message understood and how is it referred to by other messages? The necessity for this shift in focus away from agent-agent relations is supported by different empirical observations about discussion groups.

First, messages are not addressed to a specific receiver but are posted to be read by anybody who shows interest in them and invests the work for accessing them (message selection time plus interpretation time). This is in striking contrast to the message sending paradigm. Second, and maybe less obvious, there is a tendency for the sender of a message to disappear in large-scale communication processes. Life and

[3] A communication protocol "is specified by a data structure with the following five fields: sender, receiver(s), language in the protocol, encoding and decoding functions, actions to be taken by the receiver(s)" [11], p. 86ff.

death of communications in a Usenet discussion group are largely independent from life and death of the individual agents participating in the discussion. Consider a typical Usenet group such as *alt.agnosticism* which started on July 1, 1998 and currently contains more than 69.000 threads with about two dozen new messages posted each day (significantly more on weekends). The independence of communication from individual agents is nicely illustrated by the fact that from the first 10 authors posting messages on the day the group started, not a single one has contributed to the discussion during the year 2002. The phenomenon of diminishing importance of the sender is also evidenced by senders which disappear behind pseudonyms or cryptic E-Mail addresses. Such participants could contribute to a discussion under more than one name/address, or a name/address might be shared by several participants.

2. *Missing message-message relations.* Another shortcoming of AOM and the message sending paradigm consists in not explicitly modeling the references that a message establishes towards other messages. The missing perspective of message-message relations causes the analysis of the communication process to take a specific turn. Agent-oriented analysis aims at describing agent-agent relations by structural or statistical means. A typical result would be a load distribution pattern in a communication graph whose nodes represent agents and whose edges correspond to messages having been exchanged between agents.

What cannot be extracted from the sender-receiver model of communication, however, is an explicit reference structure of a communication process relating messages to other messages. Because speech acts in DAI have not been introduced to refer to speech acts explicitly, message-to-message relations outside communication controlled by interaction protocols can only be established heuristically.[4] Thus, a speech act primitive like "reject" that has been sent from agent *B* to agent *A* may be interpreted as a response referring to a "propose" previously sent from *A* to *B*. In an encounter of only two agents taking turns respectively, communicative acts indeed refer to each other according to the sequential flow of messages. In case of more complex communicative settings, however, taking temporal sequences for referential linkages is highly implausible. Whenever an agent exchanges messages with many other agents, perhaps along different protocols, or two agents exchange large numbers of messages asynchronously, not to speak of discussion forums with many participants addressing each other concurrently, heuristic referencing can no longer be considered to be reliable.

3. *High modeling complexity.* The figures behind Usenet discussion groups – a total of 700.000.000 messages with an annual increase of currently about 150.000.000 messages – render the task of modeling communication processes to be a prime challenge from the empirical as well as from the technical point of view. Huge amounts of data must be handled, a task impossible without computational assistance. More important, the data supporting the model at a given level of detail must be available. This poses a problem for AOM since, in general, nothing specific is known about the cognitive, motivational, and emotional state of the author who posts a message. To put it bluntly, there is simply no data available for modeling discussion groups on a very large scale in terms of goal-directed communication behavior of individual agents. Even if such data were available, the technical challenge of simultaneously

[4] In DAI, this difficulty seems to support "a view of the space of agent's interaction as merely the space of communication, ... where interaction histories simply result from the chaotic interleaving of the observable behaviours of single agents" [5], p 250).

running a very large number of agents (> 10.000) remains unresolved. Such a requirement is far beyond the capabilities of present agent platform technology[5].

To sum up, we have identified three major deficits of AOM with respect to the modeling of large-scale communication processes. Most important is the observation that the continuity and outcome of communication in discussion groups cannot be explained as being warranted by agents continuously participating throughout the entire process. In contrast to what normally would be expected from cooperation in multi-agent systems with persistent agents on one hand and transient entities called speech acts on the other, agents appear to be transient in forums and discussion groups, whereas messages appear to be persistently available while the discussion goes on. It is the continuous availability of messages rather than the continuous presence of identifiable agents, that keeps the communication process alive and shapes its outcomes.

3 The Alternative: Communication-Oriented Modeling

There is no doubt that agent-oriented modeling and speech-act theory have their own merits. However, with regard to analyzing and modeling complex social processes and structures as networks of communication in discussion groups, it has been shown that AOM exposes certain shortcomings and deficiencies. Hence we suggest that a different approach should be adopted: *communication-oriented modeling* (COM). Instead of modeling agent-to-agent relations we focus on modeling message-to-message relations as a methodological alternative to AOM. In our approach it is no longer the agent, but the communication event which is taken as the unit of analysis and design. COM has its conceptual foundations in a theory of communication which will be outlined in this section. This theory is, in turn, inspired by ideas taken from socionics [17], from social theories of symbolic interaction and pragmatist semiotics [19, 21, 20] and from Luhmann's sociological theory of social systems [17, 25]. It is based on three fundamental distinctions: inception and reception, observability and unobservability, and persistence and transience.

3.1 Conceptual Distinctions of a Theory of Communication Networks

1. *Reception and inception.* In our theoretical approach, communication is conceived of as a social process of messages linking or coupling to one another by *referencing*. Sociologically speaking, referencing is composed of two basic communicative operations called *reception* (understanding a message) and *inception* (producing a message). Whenever a message visibly refers to another one, this is invisibly enacted by two subsequent operations: a predecessor message is received or understood and a successor message is inceived or produced. More formally speaking, we propose to define the term referencing as a temporal event: the moment when an edge is installed between two nodes, where the nodes are two messages while the edge is made up from a pair of two complementary operations, namely reception and inception. Thus,

[5] On a platform such as JADE some hundreds of complex agent can run simultaneously, with the widely used FIFA-OS this number is one order of magnitude smaller.

in our theoretical approach to COM and to social webs of communication, the unit of analysis and design is composed of two messages and two communicative operations.

Table 2. Unit of analysis in communication-oriented modeling

Message1	Message Sign	persistent, observable
Message2	Message Sign	persistent, observable
Referencing	Reception, Inception	transient, unobservable

2. *Observability and unobservability.* Any message can be seen from two perspectives: as a physical representation of an inception or as physical representation of a reception. A message, according to our theory of communication, is an empirically perceivable object. However, it is an object of a very special kind, namely a sign-object. Being a communicative sign, it designates to non-physical, meaningful, invisible communicative operations. In a very general sense, any message (gesture, spoken word, written text, picture, icon) must be construed as the empirical manifestation of unperceivable communicative operations. In contrast to empirically visible message signs, reception and inception are black boxed. Being operations that process meaning, they are unobservable. However, they can be reconstructed from relational constellations among message signs. Reception and inception must be distinguished by an observer who establishes a meaningful relationship between different message signs by referencing. Referencing is constituted as a meaningful relationship between sign-objects via reception and inception. In any given process of communication, receptions are linked to previous and successive inceptions, inceptions to previous and successive receptions, and so on, and as the process continues, a social network of communication is dynamically formed. It is the observer's or designer's task to open the black box and to explain how a communication network is enacted by its elementary operations.

3. *Persistence and transience.* Let us assume that communicative operations are transient. Communications come and go, one operation is followed by the next, inception is coupled to reception, reception to inception, and new messages are constantly added in a continuous process of social reproduction. Communications are elementary events, i.e. discrete, temporal elements in an ever evolving network of communication. Being events, communicative operations are temporally defined by the amount of completion time they need to process the meaning of a message. Being operational elements, inception and reception take exactly the amount of time they need to create or understand a message, e.g. read a book, utter a sentence, understand a question, write a letter. In an oral conversation, but also in Internet discussion groups, operations usually appear to be very short, ephemeral events. In a scientific discourse they tend to be much longer. In any case, they simply last as long as it takes to write or read a paper. In both cases, however, in everyday life encounters of oral communication as well as in scientific discourses, any reception and any inception is processed discretely as a unique event which disappears when it is over. Of course, any operation may leave traces in memories and messages. But traces are traces of past events, not events in operation. In contrast to transient operations, messages are persistent objects, at least in textual and electronic communication. However, messages should not be misconceived as the immutable mobiles of social structure. Accordingly, a network of communication should neither be misconceived as having a static architec-

ture. There is nothing static about social structures – in COM just as in real societies – since they are dynamically reproduced by new communicative events being added and old ones being deleted. A message's persistence or transience depends, in the first place, on its physical properties. What is more interesting, however, is that messages are activated, deactivated, or reactivated by selective referencing. This is what we call a message's relative social persistence or relative transience respectively. With regard to its social persistence, a message's social visibility can be enhanced and its life-span can be extended by being repeatedly referred to in subsequent messages. The more successor messages refer to a predecessor message, the higher its social relevance and significance, and hence its social visibility. And vice versa: any message may be socially deleted as a transient social object if it is constantly ignored or non-referenced, although it may continue to persist physically.

In the course of continuous referencing a social communication storage will be built up as an unintended (or emergent) infrastructure of communication. Unless it will be drawn on again and again by subsequent messages, sooner or later any given message or communication thread will be socially forgotten. Only think of Google's millions of inactive threads which have definitively disappeared from the social process of communication, although they are still "there". Note that a social network's survival does not depend on the survival of a single messages, however. As follows from the transient character of its operations and the socially constructed persistence of its messages, a social network's continuous reproduction also depends on the disappearance of operations and related messages, on selective referencing. A social network of communication, its structural persistence (and evolution) depends on the transient character of its operational elements. These have to be permanently activated in order to produce and reproduce masses of message-objects on a large scale basis in order to provide the network with sufficient redundancy which serves, in turn, as a prerequisite for evolutionary selectivity. Operations and messages must be continuously replaced by new ones to keep the social network alive. From a functional perspective, social reproduction (as well as innovation) seems to be in need of a permanent influx of new messages, although many of these, if not most of them, will never be reactivated or re-referenced again. Hence, sociality is run like a self-referential process: Messages selectively refer to other messages and in doing so they do not only permanently reproduce the operational elements of communication (reception, inception) and their empirical equivalents (sign objects), but they shape society as a dynamic structure which is both capable of stabilization and change.

3.2 Reception + Inception = Referencing

Viewing a message as a double manifestation of reception and inception is supported by empirical evidence about navigation behavior in Internet discussion forums. Usually, what we can see in the browser, are messages referring (explicitly or implicitly) to other messages, but not human participants referring to other human participants. Communicative events are not presented in an agent-oriented manner but in a problem- or rather argumentation-oriented manner. The way in which message threads usually are visually presented seems to indicate that something like a "gestalt switch" is taking place: from agents (and cognitive processing) to messages (and communicative processing). This does not necessarily mean to radically abstract from any idea of agent or agency. However, to make progress in COM, the agent's profile and persistency must be deliberately transformed into a background feature, while the message has to be switched from background to figure. Thus, the agent is no longer presented as the predominant figure and principal attractor of theoretical attention and design strategies. Or to put it differently: Only the message level is visible or empirically accessible while the level of operations is black boxed. Operations like reception and inception must be hypothetically or theoretically disclosed.

In line with such a "gestalt switch" it should be possible to reduce the amount of cognitive assumptions needed for COM to a minimum and to draw attention from psychological or cognitive processes to designing communication socionically. It can be observed that most participants in Usenet discussion groups are actually silent most of the time. At the moment of posting a new message, a participant may explicitly or implicitly refer to one or more previous messages. In posting his or her message, he or she connects two distinctive communicative operations: a reception (of a predecessor message) with an inception (of a successor message). However, note that receptions do not automatically trigger inceptions within an agent's cognitive apparatus. In the contrary, any forward connection or junction from a reception to an inception is a highly contingent event if we assume agent autonomy. A message may be received, but the receiver may not be inclined to inceive a new message. This may happen at any point in time. As a matter of fact, any encounter and any episode of interaction sooner or later comes to an end. As long as communication is faithfully recorded and physically stored, however, any reference structure for any given message can be reconstructed from past events, since any inception must have been triggered (incidentally or causally) by a previous reception. There is nothing like a first communication event.

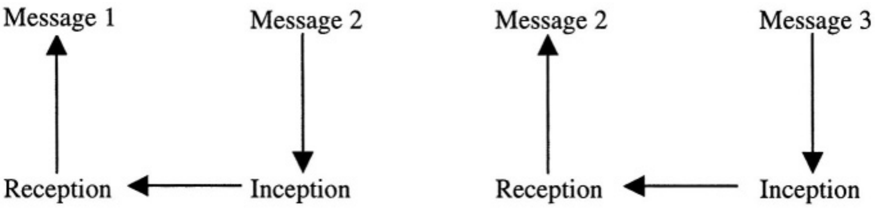


Fig. 1. Ontological dependencies in communication-oriented modeling

To sum up our theoretical assumptions in a more stringent and formal way, we can say that, in Fig. 1, the arrows do not present causal relationships but stand for formal-ontological dependencies. These can be read in an upstream fashion or against the temporal flow of communication as follows: If X exists, then Y must exist. Whenever a message exists, it can be concluded that the inception which has created the message, must have existed too, and furthermore, also the previous receptions presupposed by the inception must have existed and, accordingly, the predecessor messages that have been received earlier. What we can observe here, is a curious gap of explanation between the inception of a message and its reception. The challenge is to precisely describe and explain the conditions under which inception is followed by reception. However, when we come to the next diagram about causal relationships in the temporal flow of communication, the theoretical problem seems to consist in the difficulty to explain why a reception is followed by an inception.

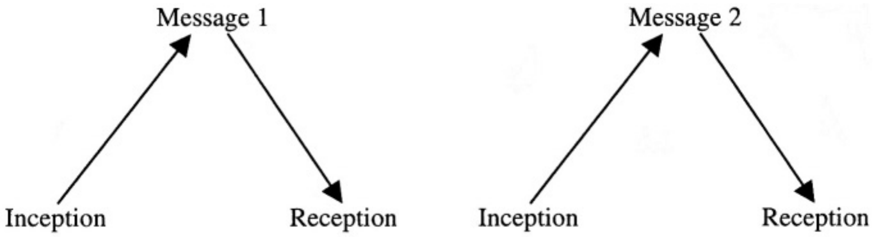


Fig. 2. Causal dependencies in communication-oriented modeling

To insist on the principle that every inception is based on a previous reception, admittedly implies to assume an actor who, for instance, gives a reply to a question. However, to point out to the fact that actors are always involved in communication and that the power to close the explanatory gap between reception and inception stems from actors and agency, misses the point in question. Viewing the issue from a functional perspective of large-scale processes of communication as in news groups, raises quite another question: What are the specific receptions an inception is drawing on and how can they be identified? Or, formulated on the level of perceivable messages: What are the particular messages from which a new message is generated or reproduced? Asking these questions means to proceed into a direction of analysis and design which is quite different from AOM. AOM is controlled by classical sender-receiver questions: Who sends what to whom? Or, in a derived form: Who are the favorite receivers of the messages coming from a sender? In contrast, to answer the questions of COM, we will have to explicate the reference structure of messages from underlying communicative operations or events. In doing so, it should be more convenient to leave the operational distinction between inception and reception behind, at least for the purposes of this paper, and to reduce both operations into a single operation: referencing. In the following section, referencing is introduced as the starting point of formalizing what we call message visibility.

4 Formalizing Message Visibility

According to our theory of communication outlined in the previous section, COM takes the relationship between a specific type of communicative events, namely the publication of messages, and the structure resulting from the references established between the messages, as its starting point. As we have just said, a publication event bundles the two complementary types of communicative operations which may be described as semantic actions that an autonomous agent is capable of performing: reception and inception. Instead of elaborating and formalizing the conceptual distinction reception/inception, it is more convenient for the purposes of this paper, to treat both operations as a single event of publishing or referencing at a higher level of abstraction. However, we will draw on the other two distinctions introduced in the previous section: observability/unobservability and persistence/transience. Hence, the formalization we propose mirrors the relationship between event and structure. It introduces two basic structures: the first describing the temporal ordering of publication events, and the second describing the reference structure of the messages. These descriptions lead us to a concept of *social* visibility which will be elaborated in this section.

4.1 Basic Structures of COM

The publication event structure (P, ≤) is a poset (partially ordered set) with the set of publication events P as ground set and the temporal ordering of events ≤ as partial order relation. Intuitively, the partial order relation p ≤ q reflects the fact that p has been published before or at the same time as q. Note that the poset structure is equivalent to making the following assumptions about the temporal ordering of publication events:

Reflexivity	$p \leq p$ for all $p \in P$
Antisymmetry	$p \leq q$ and $q \leq p$ implies $p = q$ for all $p, q \in P$
Transitivity	$p \leq q$ and $q \leq r$ implies $p \leq r$ for all $p, q, r \in P$

In an application domain where a global synchronization mechanism exists which provides a unique time stamp for each publication event, the partial order becomes a linear order, that is, the following additional property holds:

Comparability for any $p, q \in P$, either $p \leq q$ or $q \leq p$

The second structure central to COM is the *message reference structure* $(M,\leftarrow)$ which consists of the ground set M of all published messages and a binary reference relation ← on M. Intuitively, $n \leftarrow m$ expresses that message m contains a reference to message n. Structural restrictions on the reference relation arise from the fact that messages are generated by publication events. This association is established by a bijection γ: $M \to P$ which maps a message m onto the publication event $p = \gamma(m)$ that generated it. Requiring γ to be bijective amounts to assume that there is no publication event that does not produce a message (surjection), and that no two different messages are generated by the same publication event (injection):

Surjection	for each $p \in P$ there is a $m \in M$ with $p = \gamma(m)$
Injection	$\gamma(m) = \gamma(n)$ implies $m = n$ for all $m, n \in M$

The references which a message establishes to other messages are restricted by the further requirement that they should be compatible with the temporal ordering of publication events. In other words, a message may refer only to messages that have already been published at the time of its publication.

Compatibility $n \leftarrow m$ implies $\gamma(n) \leq \gamma(m)$ for all $m, n \in M$

Let us briefly discuss the implications of having messages inherit the temporal structure from publication events. For that purpose, we consider both, the publication event structure and the message reference structure, as digraphs (directed graphs) whose nodes are formed by the elements of the ground set and whose edges correspond to pairs of nodes linked by the temporal ordering or the reference relation respectively. Since the publication event structure $(P, \leq)$ is a poset, the publication event graph does not contain any cycles. This property of being a DAG (directed acyclic graph) is inherited by the message reference graph because the bijection γ induces a subgraph isomorphism which embeds the message reference graph into the publication event graph. See Fig. 3 for an example of how the two graphs relate. In order to reduce visual complexity, not the publication event graph itself but its Hasse diagram[6] has been depicted. Note that in general the message reference structure does not inherit the properties of reflexivity and transitivity from the publication event structure.

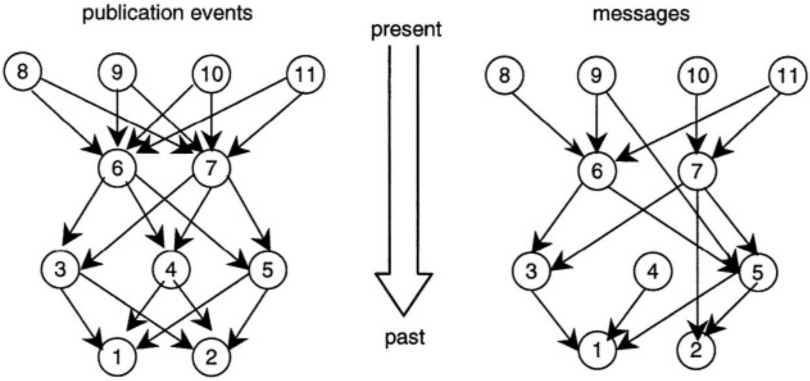


Fig. 3. Publication event graph (Hasse diagram) and message reference graph

In view of the consequence of acyclicity one might raise the question whether compatibility is not too strong as requirement. After all, web sites frequently establish cyclical links between documents and scientific publications do list some of their references as being “in print”. With regard to this issue, we must clearly distinguish between the COM framework with its inherent assumption of acyclicity on the one side, and the way that this framework is used for modeling a particular application domain such as, for instance, web pages or scientific publications, on the other side. There is no doubt that in some domains, cyclical references or references to future publication events are useful or even necessary. However, from the perspective of COM such references appear as being composed out of a sequence of references which are acyclic and directed to the past only. If a scientific paper *A* cites another paper *B* as being “in print” then some preprint version *B’* of this paper must have been published before and known by the author of *A*. It is to this already published paper *B*’ that the reference is established. The published form *B* of *B’* may now contain a discussion of related work with a cyclic reference to paper *A*. But this implies that the

[6] To obtain the publication event graph from its Hasse diagram one must (1) add reflexive edges of the type $p \leq p$ at all nodes, and (2) add the transitive closure of all edges.

cyclic reference is constructed only in retrospective by an operation identifying *B* with *B'*. To sum up, we do not see empirical reasons to abandon the requirement of compatibility which is in agreement with our socionic theory of communication.

4.2 Measuring the Social Visibility of Messages

Communicative events (in our case: publishing events) do not persist over time although they leave a persistent trace in form of the messages they generate. A closer look reveals that the distinction is one of degree rather than principle: publishing does not occur instantaneously and messages do not exist forever in the sense that they do not remain eternally accessible for references from other messages. The COM approach claims that the empirical fact of temporally limited access to messages is not caused by the technical problem of making data objects physically persistent – an issue which is studied in the context of databases and digital libraries. Even with physically or technically persistent message signs, a decrease in accessibility will occur over time because the access to a message is linked to its social visibility in the communication process. The tendency of messages to become less visible over time is counterbalanced by the tendency of references to increase the social visibility of the message that is referred.

How exactly the temporal ordering of messages and their reference structure determine the visibility of a message depends on the specific application domain and the type of communicative process that is observed. We expect to find different measures of visibility for, say, Usenet discussion groups on agnosticism and on Java Server Pages, and – outside the Internet world – for scientific publications in computer science and sociology. For this reason we cannot but give one example among the many measures possible within our framework. It needs a few definitions which are likely to be relevant for other measures of social visibility too.

Let *M* be a message reference graph. We write S_m for the set of successors of a message *m* in *M*[7]. In cases where the connection to socionic communication theory needs to be made explicit, we call S_m the *receptum of m.* The set of direct successors of *m* is denoted by DS_m. Analogously, we write P_m for the set of predecessors of *m,* also called the *inceptum of m,* and DP_m for the set of direct predecessors of *m.* Note that S_m and DS_m do not change over time whereas P_m as well as DP_m may be increased by new elements as new messages arrive which establish references to *m.* With most visibility measures, either increasing DP_m or increasing the inceptum P_m, i.e. the number of messages referring to *m,* results in an increased visibility of *m.* Note however, that also the receptum may play a role in determining visibility in some communication processes (e.g. citing outdated and exotic literature is not likely to increase scientific visibility).

Regarding the tendency for a decrease of visibility with time we only discuss the simplest case in which the publication events are ordered linearly. Furthermore, we

[7] The definition uses the graph-theoretical notion of successor (and predecessor). Node 7 in graph *M* of Fig.3 has three successors, namely nodes 2, 3, and 5, as well as two predecessors, the nodes 10 and 11. This is not to be confused with the intended semantics of the reference relation. In an analysis of literary communication where the messages represent novels and the references are given by shared stylistic elements, novels 2, 3, and 5 may well be viewed as stylistic precursors of novel 7 although, graph-theoretically, they are successors.

assume that each publication event, and, as a consequence, each message m is associated with a real number $t(m) > 0$ that serves as its time stamp. Any monotonous decreasing function is a potential candidate for describing the decay of social visibility. A nearby choice for a function measuring the *recency of a message* is $r(m) = e^{-t(m)}$ which assumes the value 1 for the present, $t(m)=0$, and exponentially decreasing but always positive values for past events. This measure is easily integrated into a measure of visibility that also takes DP_m into account:

$$visibility(m) = \sum_{n \in \{m\} \mathrm{Y} DP_m} e^{-t(n)}$$

Note that this is just one out of many possible visibility functions which abstracts, for instance, from any influence that S_m could have on visibility. Fig. 4 describes the incremental construction of the message reference graph from Fig. 3. Four cycles of a simulation are shown. In cycle 1 two messages 1 and 2 are generated and assigned the visibility value 1.0 for new messages. The number of newly generated messages at a cycle c is a random variable N(c) whose probability distribution is one of the parameters of the simulation model. We assume N(c) to be equally distributed among values from the integer range [1...4]. In cycle 2, three more messages (3, 4, and 5) are generated. These messages establish references to the already existing messages (1, 2). The number of references that a newly generated message n establishes is again a random variable $R(n)$ whose probability distribution is another parameter of the simulation model. In this case, an equal distribution among values of an integer range [1...3] was chosen. Visibility enters the play when it comes to determining the old messages that the references are directed to.

Messages with high visibility are more likely to be referenced by new messages. The exact form of the distribution $V(o)$ describing the probability with which an old message o is referenced by a newly generated message is yet another parameter of the simulation model. However, the distribution must satisfy the following consistency condition for any two old messages o_1 and o_2.

Visibility $\quad visibility(o_1) \leq visibility(o_2)$ **implies** $V(o_1) \leq V(o_2)$

In our case this is achieved by defining $V(o) := visibility(o)/total\text{-}visibility$ where *total-visibility* denotes the sum of the visibility values of all old messages. After the references have been established from the new messages (3, 4, and 5) to the old messages (1, 2), the visibility values are updated. According to the visibility function, newly generated messages are assigned the visibility value 1. For the old messages, visibility is diminished by temporal decay and increased by all incoming references. The original visibility of message 1, for instance, has decreased to $e^{-1} = 0.4$ to which adds increase of visibility by 3.0 due to the three incoming references from messages with visibility 1.0.

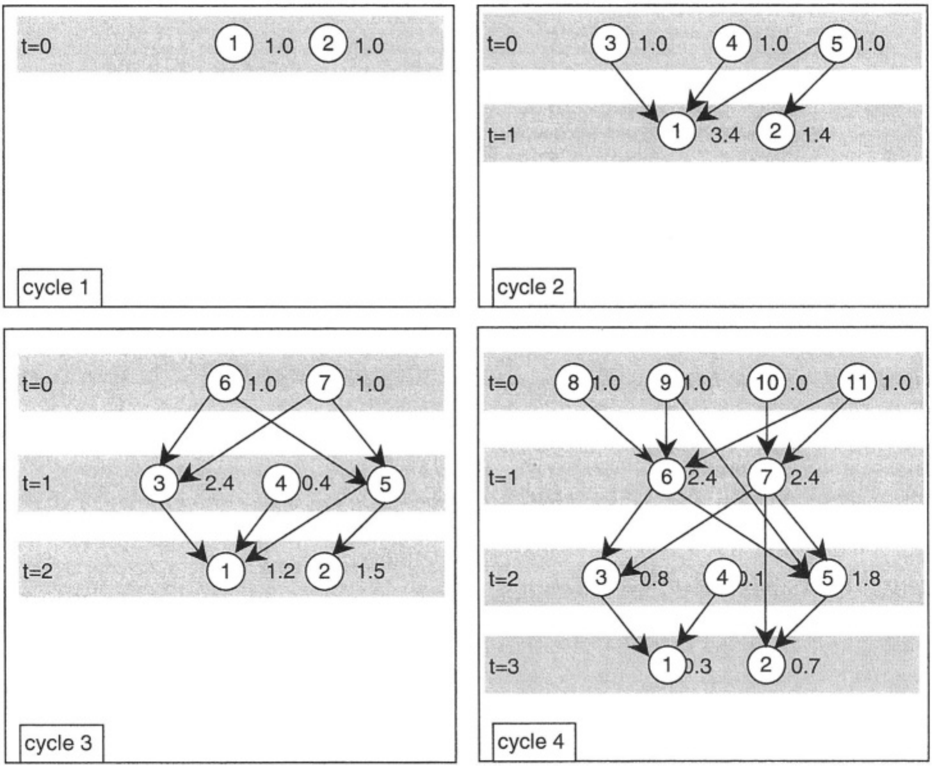
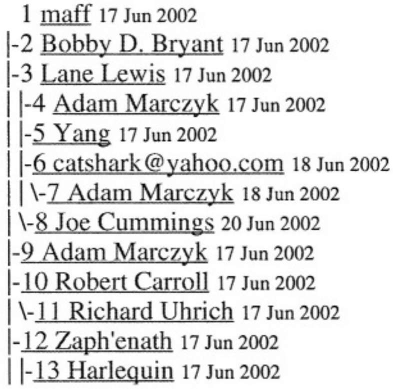


Fig. 4. Visibility values during incremental construction of a message reference graph

4.3 A Usenet Scenario

We close this section on the formal framework with an illustrative description of the constituents of a COM for a Usenet discussion process. Without detailed empirical analysis, it is, of course, not possible to come up with a fully fledged model that fits some set of empirical data. In particular, the probability distributions N(c), $R(n)$ and $V(o)$ can only be determined with respect to a concrete communication process. The first step in COM consists in identifying the messages and the reference relation in the domain. A straightforward – but not the only possible – choice regarding messages is to adopt a realistic stance: each message posted in a discussion group appears as a *message* in the model. Similarly, the reference relation in the model represents, in the simplest modeling approach, *direct references* between messages in the discussion group. These are the references established by the author's decision to post the message within a particular thread of a discussion group, generally as an answer to some other message.

Consider someone who wants to share a new type of argument refuting Creationism. This involves making a selection from an enormous range of possibilities. Among 100.000 Usenet discussion groups, the author opts for *alt.agnosticism* as the best place for publishing his argument. But this still leaves him with deciding which of the 69.000 threads to contribute to. He selects the thread *"15 Answers to Creationist Nonsense"* which contains more than 1.000 messages and decides to post his message as a comment to a message which already has produced some sequel messages, hoping that this will draw more attention to his message. Fig. 5 illustrates the tree structure of explicit references in Usenet discussion groups by showing the beginnings of the thread *"15 Answers to Creationist Nonsense"* in the discussion group *alt.agnosticism.*



```
 1 maff 17 Jun 2002
|-2 Bobby D. Bryant 17 Jun 2002
|-3 Lane Lewis 17 Jun 2002
| |-4 Adam Marczyk 17 Jun 2002
| |-5 Yang 17 Jun 2002
| |-6 catshark@yahoo.com 18 Jun 2002
| | \-7 Adam Marczyk 18 Jun 2002
| \-8 Joe Cummings 20 Jun 2002
|-9 Adam Marczyk 17 Jun 2002
|-10 Robert Carroll 17 Jun 2002
| \-11 Richard Uhrich 17 Jun 2002
|-12 Zaph'enath 17 Jun 2002
| |-13 Harlequin 17 Jun 2002
```

Fig. 5. Threads in a Usenet discussion group

COM captures what is essential here: messages are published rather than sent. The crucial point is not that messages are made accessible to anyone who is interested but that COM takes into account the fact that the amount of interest that any person can invest is highly limited, forcing everyone to make choices.

In a complex environment with thousands of messages arriving in a relatively short period of time, there is a limit to cognitive orientation. In other words, large-scale discussion processes exert a tremendous pressure on the cognitive capacities of human actors. Unlike electronic mail which is easy to handle since it operates on a one-to-one base, it is extremely demanding to keep up with a communication process which is inherently open and runs on parallel forums and threads. Participation in Usenet discussion groups requires an amount of selective attention from the user which far exceeds the limits of ordinary human capabilities. A user who wants to know what is discussed, whether there are parallel discussion groups on similar topics going on at the same time in different places, and when, where and how to post his own statement, is in permanent danger of getting lost. To compensate for limited cognitive capacities, COM could prove helpful since it suggests to reduce the complexity of Internet discussions in a way similar to what social networks do when constructing variable reference structures from an ongoing process of communication .

5 Future Work

As can be seen from the previous sections, there is still much work to do to develop COM as a powerful methodology and to demonstrate its scientific fruitfulness. In this section, we shall at first give an idea of how our approach to communication could beneficially be applied to current work in the multi-agent field. Secondly, we shall briefly address some more theoretical issues which seem to be particularly promising with regard to modeling and simulating artificial societies.

Interesting enough, not only an actor's cognitive orientation could be fostered by introducing methods of complexity reduction at the level of group interaction. What is more, strategies and techniques of complexity reduction and coherence management are of even higher relevance at the social level of large-scale communication. Take, for instance, the case of redundancy in parallel forums: Usually, it will be unfeasible for a moderator to prevent people from talking about the same issue in different threads or sub-forums at the same time. And even if so: Would an intervention be beneficial to the participants anyway? Who, in the role of a moderator, should be able to eliminate a posting as off-topic, or, in case of acceptance, who should decide where to insert which statement at which point in time or whether a contribution belongs to a specific topic or not? There is no global knowledge or privileged overview in a dynamic web of communication. At the global level there is nothing like a point of observation, whence everything could be seen as it really is. *Any observational perspective is part of the process and is, therefore, limited.* Consequently, "problem solving" in open forums like Usenet groups with mass participation cannot be viewed as being a pre-structured by centralized coordination to such an extent that it leaves but some of the details to be settled by spontaneous contributions at the local level. In the contrary, the entire process is a contingent phenomenon. Any collective decision, any shared common view, consent or dissent, is but the outcome of communications referring to other communications. A software tool designed to assist participant users and moderators has to be capable to observe and reconstruct a discussion in terms of messages referring to other messages. It should be able to register and monitor the flow of contributions, to analyze the process of new messages linking to previous ones, and to visualize temporal and referential patterns of communication. Additionally, an assistant tool should provide technical means of analyzing and simulating processes with very large numbers of communication events to study the effects of different scenarios on pattern formation and structuration. In the hands of a user, this should be an efficient instrument for individual orientation.

Moreover, in a technically more complex version, COM could be applied on top of a multi-agent platform. Consider a team of moderator-agents coordinating a large-scale debate in the Internet: How could these moderators profit from COM? What we suggest is to develop tools for distributed moderation. Obviously, moderating a discussion forum is more demanding than just preparing an individual statement and selecting the right moment for intervention. In order to come to grips with a communication process of ever growing complexity and redundancy, with different topics and opinions discussed in parallel, with multiple threads of argumentation, perhaps with sharp lines of conflict and dissent, there is a need for distributed coordination. To be successful, coordination should not try to control the debate but take advantage of *evolutionary tendencies towards differentiation or redifferentiation* as they occur in a quasi natural flow of communication. Hence, it might be helpful to distribute the moderator's task among several agents. Designing moderation as a cooperative task of coordination, begins with aggregating (or disaggregating) tasks and assigning each moderator agent with a specific responsibility. Instead of viewing the entire process with equal attention, one moderator agent might specialize on a specific feature, e.g. on conflict mediation, another focuses on a specific topic, a third one follows up a different topic, a forth one analyses temporal patterns, turmoils and "hot spots" etc. To do so, all agents work with a COM-based communication analysis. However, they do so selectively and with different attentions. Again, this follows from the fact that *any observational perspective is limited.* Observing conflicts, for instance, means to distinguish between consent and dissent, irrespective of other features of the debate; temporality observation operates with the distinction immediate/postponed responses; topics observation distinguishes between on/off topic etc. Multi-agent moderation means that all these different observational distinctions will have to be cooperatively integrated, and this would be the task of an agent platform taking COM-based communication analysis as its input.

Coming to our theoretical perspectives, there are two issues that deserve particular attention in future work: social visibility, empirical observation, differentiation of sub-forums and reflexive or meta communication. These issues are closely interrelated and they just have been touched on in this section with regard to moderating discussion groups with COM and multi-agent systems.

1. *Visibility and observation.* A credible concept of social visibility should be based on the assumption that any observational perspective is limited and that there are always several perspectives installed. So far, we have considered social visibility as a global concept that applies to all publication events in the same way. This is also plausible for a small-scale process, e.g. a single thread in a discussion forum. In that case it can be assumed that all agents involved in publishing share – up to some tolerable amount of error – the same measure of social visibility. However, in large-scale systems different perspectives of observation, hence different measures of visibility will coexist. In particular, the visibility of a single reference which forms the recursive base of the computation of the visibility function may differ. Consider a HTML document whose references, i.e. the hyperlinks, are visible only to software such as a web browser which can interpret HTML. The same holds for a human reader confronted with documents in different languages. Only some readers will be able to extract references from a text written in Japanese kanji characters. Therefore, an obvious refinement of our approach aims at allowing different publication events to work with different notions of social visibility. Another refinement concerns the computation of visibility itself. It is quite possible that forthcoming experience with modeling communication processes will show that other factors beside the temporal ordering of messages and their reference structure affect visibility. Obvious candidates are identifying operations on messages such as the one mentioned above which establishes the identity of content between different versions of a message. This is by no means a trivial problem. The fact that a message is part of a group of "identical" messages, e.g. different versions of the US constitution, contributes significantly to its visibility. In the same way, the effect of the author of a message on the message's social visibility can be modeled as the effect of an identifying operation which groups messages by authors.

2. *Differentiation and reflexive communication.* We must try to develop models of social differentiation and re-differentiation to describe and simulate contrasting modes of dynamic reproduction, e.g. stable reproduction and evolutionary change of social

structures in complex networks of communication. To achieve this goal, particular attention must be paid to the notion of temporality as expressed by the event structure of communication. With regard to social visibility values in the incremental construction of a message reference graph (Fig. 4), we can make the assumption that, for instance, a value lower than 0.4 means that a message falls under a given threshold of social visibility. Taking invisibility into account as an empirical phenomenon which is socially constructed in everyday communication by ignoring (non-referencing) a message, we need fine-grained patterns of referencing (adoption, rejection) recurring to the operational distinction between reception and inception again. From here, it should be possible to model and simulate social evolution: By selective referencing, a communication network gradually begins to separate into two (or more) different forums of communication, perhaps organized around different topics, opinions, authorities, authors etc. – different forums which will eventually have no longer any access to each other although they have been generated from the same home-forum. This means that we will have to study social invisibilization at a more specific level of investigation. Moreover, we need theoretical models of what sociologists would call the innate reflexivity of social communication. A discussion becomes reflexive when participants begin to communicate about the discussion as a communication process. Any communicated observation concerning the discussion at a meta level (rules, outcomes, topics, and standard) inevitably inserts reflexivity into the process. Obviously, reflexive communication may have an enormous impact on the entire process of communication, and therefore it must be regarded as one of the central tenets that we will have to cope with in future work.

6 Related Work

Our work is related to different fields of research in DAI, sociology, and socionics. To begin with socionics [18, 12], our approach is closely related to expectation-oriented analysis and design, based on modeling expectation by means of a social mirror [3, 14]. This work is, similar to ours, inspired by ideas taken from social systems theory [17]. In contrast to our proposal, however, the social mirror is still based on speech acts. Moreover, it does not explicitly allow to model the temporal structures of social change in terms of complex chains enacted between communication events.

Coming to DAI research, our approach has been encouraged by recent publications suggesting to put communication and interaction on top of the agenda for multi-agent systems design [28]. Our proposal differs from DAI's paradigm of communication dominated by KQML speech act primitives and FIPA standards. Communication in agent-oriented modeling has its focus on interoperability issues such as the common language problem [13], conversation-type of interaction [2], and dialogue-oriented communication in small groups of agents [4]. In our paper, we take interoperability for granted. Rather that dealing with interoperability issues, we want to contribute to the coordination problem of DAI. As has been recently suggested [5], agent interaction should not be viewed as merely occurring within a given technical infrastructure of communication. In order to deal with the complexity of interagent communication more efficiently, theories and tools are needed to design coordination into multi-agent systems via social rules and collective commitments. In our paper we do not directly tackle the coordination problem by providing, for instance, agent ensembles with off-line designed social laws [23] or models of other agents' beliefs, abilities, and preferences [10]. We do so rather indirectly by studying a very simple on-line mechanism of a highly abstracted social structure that emerges as a temporal pattern of communicative events.

In fact, this is the central tenet of our paper, and it is clearly inspired by the sociological turn from action to social interaction and communication [19, 17, 25]. According to Luhmann's theory of autopoietic social systems, communication must be construed as the temporal element (or basic operational event) of social systems that reproduce themselves by permanently producing the very elements of communication they are made up of. Our proposal to represent social evolution as a combination of a reference structure and an event structure, is different in that it has been rather freely adopted from Mead‘s and Luhmann's views of society as a dynamically evolving network of communication. Moreover, our approach to COM has also been cross-fertilized with other sociological concepts of communication taken from conversation analysis, and objective hermeneutics [9, 22]. Our distinction between perceptible messages and unperceptible albeit meaningful and hence accessible communicative operations is taken from there.

Last not least, our work is related to the study of Usenet discussion groups and other Internet forums with regard to democratic participation in large-scale public debates [16] and to methods of social network analysis [26] applied to internet discussions [1, 15]. A major deficiency of social network analysis (SNA) must be seen in the fact that it is based on an agent-oriented and rather static methodology. Indeed, SNA seems to suffer from a considerable lack of providing appropriate means to describe the evolutionary dynamics of social networks. Again, our own approach to COM is specifically designed to analyze and simulate evolutionary processes of network configuration and might, perhaps, contribute some ideas to render SNA more dynamic.

References

1. Albrecht, S.: Structuring Large-Scale Online Debates - Making Use of Network Analysis Methods. Paper presented at the Sunbelt XXI International Social Network Conference, Budapest, Hungary, April 25–28 (2001)
2. Barbuceanu, M., Fox, M.S.: The Design of a Coordination Language for Multi-Agent Systems. In: Müller, J.P., Wooldridge, M.J., Jennings, N.R. (eds.): Intelligent Agents III. Agent Theories, Architectures, and Languages. ECAI '96 Workshop (ATAL) Budapest, Hungary, August 1996 Proceedings., Springer-Verlag Berlin, Heidelberg (1997) p. 342–356
3. Brauer, W., Nickles, M., Rovatsos, M., Weiß, G., Lorentzen, K.F.: Expectation-Oriented Analysis and Design. In: Wooldridge, M.J., Weiß, G., Ciancarini, P. (eds.): Agent-Oriented Software Engineering II. Second International Workshop, AOSE 2001, Montreal, Canada, May 29, Lecture Notes in Computer Science, Vol. 2222. Springer-Verlag, Berlin Heidelberg New York (2002) p. 226–234
4. Bretier, Ph., Sadek, D.: A Rational Agent as the Kernel of a Cooperative Spoken Dialogue System: Implementing a Logical Theory of Interaction. In: Müller, J.P.,. Wooldridge, M.J., Jennings, N.R. (eds.): Intelligent Agents III. Agent Theories, Architectures, and Languages. ECAI '96 Workshop (ATAL) Budapest, Hungary, August 1996 Proceedings, Springer-Verlag Berlin, Heidelberg (1997) p. 189–204

5. Ciancarini, P., Omicini, A. and Zambonelli, F., “Multiagent System Engineering: The Coordination Viewpoint”. In: Jennings, N.R; Lespérance, Y. (eds.): Intelligent Agents VI: agent theories, architectures, and languages; proceedings / ATAL’99, Orlando, Florida, USA, July 15 – 17, 1999- Berlin; Heidelberg; New York: Springer, (2000) p. 250–259
6. Comer, D.: Computer Networks and Internets. Prentice-Hall: Upper Saddle River, NJ. (2001)
7. Finin, T. W., Fritzson, R., McKay, D. and McEntire R.. KQML as an agent communication language. In: Proceedings of the 3rd International Conference on Information and Knowledge Management (CIKM ’94), Gaithersburg, Maryland, ACM Press (1994) p. 456–463
8. FIPA: Foundation for Intelligent Physical Agents. FIPA ’99. http://www.fipa.org.
9. Garz, D. und Kraimer, K.: Die Welt als Text. Zum Projekt einer hermeneutisch-rekonstruktiven Sozialwissenschaft. In: Garz, D. und Kraimer, K.: Die Welt als Text. Theorie, Kritik und Praxis der objektiven Hermeneutik. Suhrkamp, Frankfurt am Main (1994) p. 7–22
10. Gmytrasiewicz, P.J. and Durfee, E.H.: Rational Communication in Multi Agent Environments In: Autonomous Agents and Multi-Agent Systems 4 (2001) p. 233–272
11. Huhns, M.N. and Stephens, L.M.: Multiagent Systems and Societies of Agents. In: Gerhard Weiss (ed.): Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence, MIT (1999) p. 79–120
12. Kron, T.: Luhmann modelliert. Ansätze zur Simulation von Kommunikationssystemen. Leske + Budrich, Opladen (2002)
13. Labrou, Y. Finin, T. and Peng, Y.: The Interoperability Problem: Bringing together Mobile Agents and Agent Communication Languages. IEEE. In: Proceedings of the Hawaii International Conference On System Sciences, January 5–8, 1999, Maui, Hawaii (1999) http://www.cs.umbc.edu/~finin/papers/hicss99.pdf
14. Lorentzen, K.F., Nickles, M.: Ordnung aus Chaos - Prolegomena zu einer Luhmann’schen Modellierung deentropisierender Strukturbildung in Multiagentensystemen. In: Kron, T., (ed.): Luhmann modelliert. Ansätze zur Simulation von Kommunikationssystemen. Leske + Budrich, Opladen (2002) p. 55–113
15. Lübcke, M., Kommunikation im Netz. Eine Analyse am Beispiel von Newsgroups. Diplomarbeit im Fach Soziologie, vorgelegt bei der Universität Hamburg (2000)
16. Luehrs, R., Malsch, T., Voss, K.: Internet, Discourses and Democracy. In: Terano, T. et al. (eds.): New Frontiers in Artificial Intelligence. Joint JSAI 2001 Workshop Post-Proceedings. Lecture Notes in Computer Science, Vol. 2253. Springer-Verlag, Berlin Heidelberg New York (2001) p. 67–74
17. Luhmann, N.: Soziale Systeme. Grundriss einer allgemeinen Theorie. Suhrkamp, Frankfurt/Main (1984)
18. Malsch, T.: Naming the Unnamable: Socionics or the Sociological Turn of/to Distributed Artificial Intelligence. In: Autonomous Agents and Multi-Agent Systems 4 (2001) 155–186
19. Mead, G.H. (1934): Mind, Self and Society. Chicago (1934)
20. Mill, U.: Technik und Zeichen. Über semiotische Aktivität im technischen Kotext. Nomos, Baden-Baden (1998)
21. Morris, C.: Signification and Significance. A Study of the Relations of Signs and Values. MIT, Cambridge, Mass. (1964)
22. Schneider, W.L.: Intersubjektivität als kommunikative Konstruktion. In: Fuchs, P. and Göbel, A. (eds), Der Mensch - das Medium der Gesellschaft. Suhrkamp, Frankfurt am Main (1994) p. 189–238
23. Shoham, J. and Tennenholtz, M. Social laws for artificial agent societies: Off-line design. In: Artificial Intelligence 73 (1995), p. 231–252
24. Singh, M.P. Agent communication languages: Rethinking the principles. IEEE Computer, 31 (12):, December 1998, p. 55–61

25. Stichweh, R.: Systems Theory as an Alternative to Action Theory? The Rise of 'Communication' as a Theoretical Option. In: Acta Sociologica 43 (2000) p. 5–13
26. Wasserman, S., Faust, K.: Social Network Analysis. Methods and Applications. Cambridge University Press, Cambridge et al. (1994)
27. Wegner, P. Why interaction is more powerful than computing. Communications of the ACM, 40 (5):, May 1997. p. 80–91
28. Wooldridge, M.J., Jennings, N.R. and Kinny, D.. A methodology for agent-oriented analysis and design. In: Proceeding of the Third International Conference on Autonomous Agents.. ACM, Seattle (WA), May, 1–5 1999. p. 69–76

Modeling Product Awareness Rates and Market Shares

Filippo Neri

University of Piemonte Orientale
Dipartimento di Scienze e Tecnologie Avanzate
Corso Borsalino 54, 15100 Alessandria AL, Italy
neri@di.unito.it

Abstract. An agent based tool for analysing consumers/markets behaviour under several rate of information diffusion is described. This methodology allows for the study of tradeoffs among several variables of information like product advertisement efforts, consumers' memory span, and passing word among friends in determining market shares. Insights gained by using this approach on an hypothetical economy are reported.

1 Introduction

The diffusion of an Internet based economy, that includes even the less valuable transactions, is day by day more evident. The existing information infrastructure has allowed the exploitation of new methods to contract the purchases of goods and services, the most notable of which is probably the agent mediated electronic commerce [10,12]. In this economy, autonomous agents become the building block for developing electronic market places or for comparing offers across several seller's websites (shopbots) [12,16,11]. The possibilities offered by the new shopping environment results in the consumer adopting a (possibly) completely new decision making process to select which product to buy among the available offers. Our aim is to use an agent-based market place to qualitatively simulate the diffusion of products' awareness across the Internet and its impact on customer choices. Another important motivation in our decision to adopt an agent based simulation framework is that we aim to study how individual history and limitations impact on group dynamics. As many commercial scenarios could be selected, we chose to model a simple commercial interaction. Different groups of consumers have to choose one product between a set of perfect substitutes that differ in price, advertised lifestyle associated with the product and the advertising effort to initially penetrate the market. Our objective is the to understand how a sequence of repeated purchases is affected by the trade off among the previous variables, the consumers' desires and limits, and the diffusion of the awareness about the existing products. The modelling ultimate goal would be to capture the common experience of choosing, for instance, among alternative brands of Italian Pasta packages displayed in the webpage or on the physical shelf of our grocery store.

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 134–144, 2004.

Our research scope is in between the investigation of consumer decision making [1,6] and the study of electronic based economies of software agents, shopbots economies for short, [12,9]. In the following we describe the relationships of our work with both fields. Consumer decision making has received attention from a number of different research fields including psychology and the quantitative modelling communities. Psychological research aims to understand the reasons underlying the decision making process, whereas quantitative modelling community uses a variety of techniques from statistics, machine learning, and software agents to quantitatively model the factors that are involved in the decision making process. A summary of the most relevant works in the area is reported.

Cognitive investigation of consumer decision making. Bettman [1] and Bettman et al. [2] investigate how consumers decide what product to buy. They propose that consumers have a limited (information) processing capability, they act in order to satisfy a need, and usually do not have a well defined set of preferences to be used in product selection. Instead they construct them using a variety of strategies which depend on the situation at hand. Bettmann's work aims to make explicit the cognitive framework (i.e. state its underlying constraints) used by the consumer to then build the mental model used when deciding what to buy. Practically, a ready-to-use model of the consumers able to provide quantitative indication cannot be immediately derived by Bettman's work. From our perspective, Bettman et al.'s work show that consumers engage in a mental process consistent with weighted adding in less emotional buying tasks (i.e. buying groceries vs. buying an house). Also they show that choice processes can be selective (some products are filtered out), comparative (among the filtered remaining products) and influence the items stored in the consumer memory. Hoyer [8] proposes that consumers used different decision making strategies, not only because of individual differences, but depending on the high/low involvement (i.e. risk and/or emotional impact of purchasing the product) they feel toward the product category. Out-of-the-store decision making strategies, meaning that the consumer has already decided which product brand she will buy before reaching the store, are considered and empirical test of these hypothesis are carried out. [20] proposes a framework to analyse the strategies used by consumers to choose between alternative products. Strategies are evaluated by volunteers, participating in a psychological experiment, in term of the easiness to be remembered and applied at the moment of purchase. No attempt to model real consumers in a real store is however address.

Quantitative analysis of consumer decision making. Guadagni and Little [6] uses a multinomial logit model of brand choice to predict the share of purchases by coffee brand and (package) size. The model has been developed by using data collected through optical scanning of products at check out in supermarkets. This work is similar to our approach but we differentiated in term of technology used (software agents vs multinomial logit model), of the explicit modelling of consumers types and preferences and because we want to take into account information exchange among consumers between repeated purchases. Currim et al. [4] show how to use decision trees to represent the process consumers use to integrate product attributes when making choices. Models of the individual consumer or of consumers segments are studied in the case of choosing among alternative brands of coffee. The work is exploratory in nature and no attempt to calculate product market shares is made. Smith and Brynjolfsson [18] shows how data collected by an internet shopbot, a web-engine that compare offers for the same product but from different retailers, can be used to analysed consumer behaviours and their sensitivity to issues like product's price, shipping conditions and brand name of the retailers. Degeratu et al.[5], for instance, explores how correlations between brand name, price and sensory attributes influence consumer choice when buying on-line or off-line. They conclude that prices for products in the same category are not the main drivers for consumer decision both when shopping on-line and off-line. Instead a combination of price together with additional information, intrinsically product dependent, is used buy the customer to take a decision. Degeratu et al.'s approach is based on a stochastic model of the consumers whose parameters are inferred by real data. In this case the consumer model is a probabilistic equation where random variables account for the variation observed in the data. Our approach, instead, aims to characterise typologies of consumer through an explicit definition of the key drivers under the buying decision.

Lynch and Ariely [9] try to understand the factors behind purchases made in a real world experiment of wine selling across different retailers' websites, but no consumer model is produced.

Software agents modelling of market environments. Rodríguez-Aguilar et al. [16] explore how to use agents to define and study (electronic) auction markets and proposed that such competitive situation constitute a challenge for software agents research in the area of agent architectures and agent based trading/negotiation principles. We agree with them and we plan to evaluate some of their idea to enrich the communication part of our simulation approach. Sophisticated interactions between agents, negotiation strategies, in market-like environments are also being studied, see for instance [15,19,17]. Our approach relies on a simple exchange of information for the moment. In term of relationships of our work with research in the area of shopbots economies, we not that some researchers take a very long term view about the ecommerce phenomena envisioning economies of shopbots [10,12,16]. For instance, Kephart et al. [10] try to model large open economies of shopbots by analysing an economy based on information filtering and diffusion towards targeted shopbots (customers). Quite differently, we try to capture the commercial phenomena in a more near future where customers are human beings with their intrinsic limit in information processing, having the need to trust the bought product and to feel supported, and reassured about their purchasing choice as their best possible choice. We share with Kephart et al. the desire to analyse and understand how the information flow can affect such economy. Indeed our aim is to use an agent-based market place to qualitatively simulate the diffusion of products' awareness across the Internet and its impact on customer choices [13].

Hales' work [7] also related to ours. Hales explores the relationship between agents, their beliefs about their environment, communication of those beliefs, and the global behaviours that emerge in a simple artificial society. Our work differs however for the domain and the focus toward defining the individual behaviour responsible for an observable macro effect at the level of the whole economy.

To further extend our work, a more sophisticated approach to modeling the electronic market place may have to be selected in order to take into account negotiation protocols or virtual organisation formation as, for instance, described in [15] or to account for additional brokering agents as describe in [19]. In the near future, we would like to investigate the emergence of information diffusion strategies by using a distributed genetic algorithm [14].

The paper is organized has follow: in section 2 a description of the market place simulation is reported, in section 3 the performed experiments are commented and, finally, some conclusions are drawn.

2 The Virtual Market Place

The architecture of the agent based virtual market place is quite simple: one purchasing round after the other, groups of consumers, modelled as software agents, select which product to buy according to their internal status. The internal status takes into account the consumers' preferences for a product and her awareness about the product's benefits and image. This process based description of the buying experience matches what most people experience when selecting among alternative wholemeal breads or milk chocolate bars at the local grocery store [1]. In the simulator we represent both products and consumers as software agents. A product is a collection of an identifier, a price, an effort to describe its features/benefits on the package, an effort to bound the product to the image of a lifestyle (brand) and an initial advertisement effort to penetrate the market. It is important to note that the scope of this work is to consider products that are substitute one for the others but differ in price or other characteristics. The idea to model products as software agents is new.

A consumer is a (software) agent operating on the market and driven in her purchases by a target price, a need for understanding the product benefits, the lifestyle conveyed by the product brand, and the initial marketing effort put into placing the product in the market. The consumer can remember only a constant number of products (memory limit) for a constant number of rounds (memory duration), and she may share with her friends her opinion about the known products. It is worthwhile to stress that the memory span limits the consumer awareness of the available products. For instance, if a consumer had a memory limit of 3, she would be aware of 3 products at most and she would make her choice only among those three products. A consumer will not remember a product, if its memory has already reached its limit, unless it is better of an already known product thus replacing it. However, round after round, consumers talk to each other and they may review their opinions about the products by updating their set of known products. Our interest lays in forecasting the product market shares (percentage of bought products) on the basis on the previous factors. In order to evaluate the feasibility of our approach, we developed from scratch a basic version of the market place simulator and performed some experiments under constrained conditions.

In the following the detailed descriptions of both the simulator's architecture and the experimental setting is described. In the simulator, each product is defined by an identifier (Id), a selling price (Price), an effort in describing its benefits on its package, an effort to convey a lifestyle (image), and an effort to initially penetrate the market. In the current version of the simulator, parameters only assume binary values.

As an instance, in the initial series of experiments, all the products prices and characteristics are selected to cover a wide range of significant offers as follow:
Product(Id, Price, Description, Image, InitialAdvertisement)
Product(0, LowValue, LowValue, LowValue, LowValue)
Product(l, LowValue, LowValue, LowValue, HighValue)
Product(2, LowValue, LowValue, HighValue, LowValue)
Product(3, LowValue, LowValue, HighValue, HighValue)
Product(4, LowValue, HighValue, LowValue, LowValue)
…
Product(14, HighValue, HighValue, HighValue, LowValue)
Product(15, HighValue, HighValue, HighValue, HighValue)
The constants 'LowValue' and 'HighValue' correspond to the values 0.2 and 0.8. The Price, Description and Image parameters are used to evaluate a customer's preference for the product, whereas the InitialAdvertisement parameter defines the initial awareness of the product among the customers. So, for instance, a product defined as Product(x, LowValue, LowValue, LowValue, LowValue) is especially targeted toward price sensitive consumers that do not care about knowing much on the product. And with an initial penetration rate of 0.2, on average, 20% of the consumers are aware of its availability at the beginning of the first buying round. Finally, it is worthwhile to note that, in the above list, odd and pair numbered products differ only because of a different initial advertising effort.

A similar representation choice has been made to represent customers. Four groups of consumers are considered. For the scope of the initial experiments, we concentrate on customers whose target product has a low price but differs in the other features. Consumer groups are represented as follows:
Customer(Price, Description, Image)
Customer(LowValue, LowValue, LowValue) (bargain hunters)
Customer(LowValue, LowValue, HighValue) (image sensitive
Customer(LowValue, HighValue, LowValue) (description sensitive
Customer(LowValue, HighValue, HighValue) (image and description sensitive
Through the selection of target values, we tried to capture the following categories of customers: the bargain hunters, the brand sensitive ones, the package sensitive ones (i.e. are interested in its nutrition values, its composition, its eco-

logical impact, etc.), and those that are both brand and package sensitive. It is important to note that each customer does not necessary known the same products than other consumers because of the individual memory and of the initial random distribution of a product awareness among consumers. During each round, a consumer chooses to buy the product that most closely matches her preferences.

According to Bettman [1] and [8], we approximate the product matching process by means of a weighted average function defined as follows:

$$\begin{array}{l}
Preference(product) = \\
(max(product.Price, target.Price) - \\
target.Price)^2 + \\
(min(product.Description, Description) - \\
target.Description)^2 + \\
(min(product.Image, target.Image) - target.Image)^2
\end{array}$$

The preferred and selected product is the one with the lowest value of the Preference function among the ones known by the customer. Alternative expressions are under study.

Also each customer does not necessarily known the same products than the others because of the different distribution of the products depending on their initial marketing effort. The reported experiments aim to understand the impacts of the following factors in determining the final product market shares: customer preference definition, initial market penetration effort, number of friends in passing the word of known products, and memory limit.

In the initial group of experiments we aimed to investigate some hypothesis on the impacts of the diffusion of product awareness and shift in the consumers' behaviours [13]. The obtained results are promising and confirm the feasibility of the approach. They are however far from being conclusive in term of hypothesis testing. Indeed in order to perform extensive and informative experiments, the virtual market place simulator should be completely re-engineered to facilitate its use and the definition of hypotheses/rules governing the consumers' behaviour.

3 Experimental Results

The goal of the experimentation is to show that our tool can capture some of the inherent complexity behind the determination of the product market shares by considering a variety of factors that impact on this economic phenomena. These factors include the customers' expectations for a product, the limited memory span and duration that consumers reserve to remember available products, and the diffusion of the product awareness among consumers by initial advertisement and further passing by word. Value ranges for this variables have been selected accordingly to past experience with consumers behaviour. All the reported experiments refer to a hypothetical economy and are based on the following basic settings. During each round, 400 consumers (one hundred for each of the four consumer types) select which of the 16 products to buy. Only products that the consumer remembers (i.e. appearing in its memory list) compete for being purchased. The economic process is repeated for 100 rounds. For each experiments, the reported figures are averaged over 3 runs.

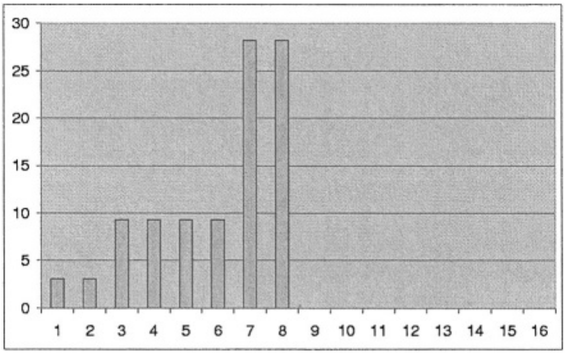


Fig. 1. Ideal market share distribution in presence of a perfect product awareness or perfect information flow.

As a baseline for evaluating the economic process, we consider the situation where each consumer is fully aware of all the available products since the first round. As all the consumers are oriented towards products with low price but with different characteristics, it is straightforward to calculate that the product market shares stay constant over the 400 rounds and correspond to the values reported in Fig. 1. In the picture, the product's identifiers appear on the x axis, and the market shares on the y axis. Thus for instance, Product 6 will achieve a 9.3% market share. It is worthwhile to note that the product from 9 to 16 have a 0% market share because, in the range from 1 to 8, there exists a product with identical features but with lower price.

If we were in this ideal situation, every consumer would be able to make the best pick among the available products. Unfortunately, in the real world, full knowledge about the available choices is not common and product awareness is the results of a variety of factors including advertisement, passing by word among friends and memory capacity. The impact of these factors on the product market shares is taken into account in the following experiments.

Let us consider the case where consumers do not have any friends or do not talk about products to friends (average number of friends or avgf =0), they can remember only 2 products at the time (memory limit or ml=2), and they remember each product for 20 rounds unless either they keep buying it or they are told about by their friends. The initial (end of round 1) and final market shares (end of round 100) appear in Fig. 2.

It appears that the initial and final market shares are very alike and that the higher the effort in penetrating the market the better the market share (compare odd and even numbered products). The market share distribution is biased toward low priced product, this is to be expected given the customers'

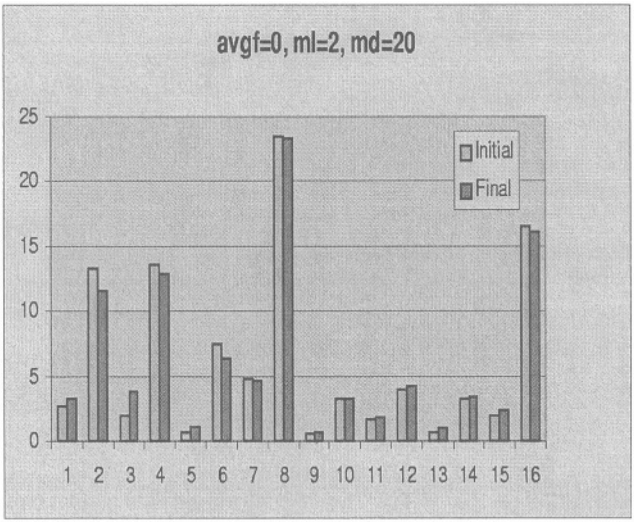


Fig. 2. Product market shares in the case of consumers not talking to their friends about their shopping (avgf=0), remembering at most 2 products (ml=2) and with memory duration of 20.

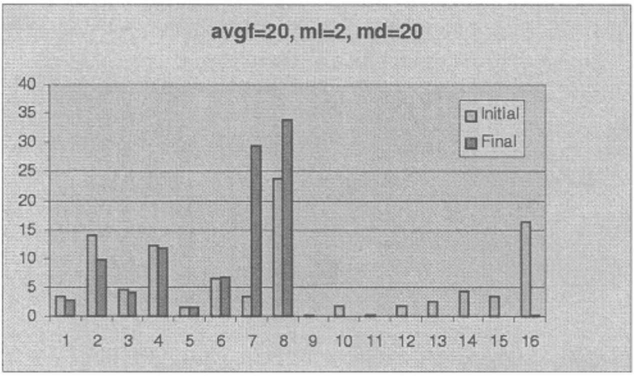


Fig. 3. Product market shares in the case of consumers talking to about 20 friends about their shopping (avgf=20), remembering at most 2 products (ml=2) and with memory duration of 20.

preferences. But, still, some high price products achieve a significant portion of market because of the limited memory span of the consumers that would prevent him to compare and choose among more alternatives.

If we alter the previous scenario just by increasing the number of friends to 20, we obtain quite a different distribution of market shares, Fig. 3.

The pattern of the initial market shares is, of course, similar to that of the previous scenario but the final shares tends to converge towards the ideal ones.

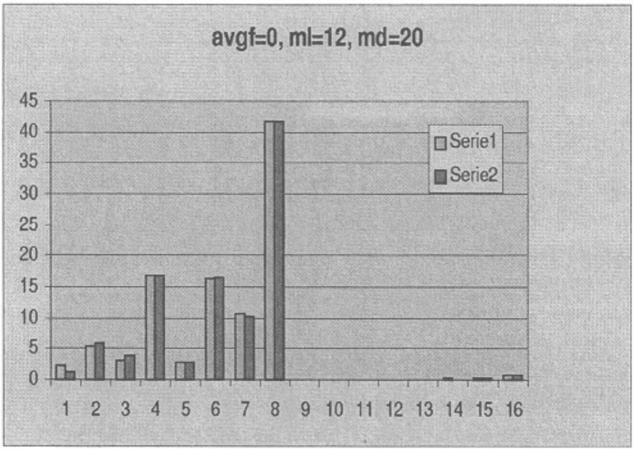


Fig. 4. Product market shares in the case of consumers not talking to their friends about their shopping (avgf=0), remembering at most 12 product (ml=12) and with memory duration of 20.

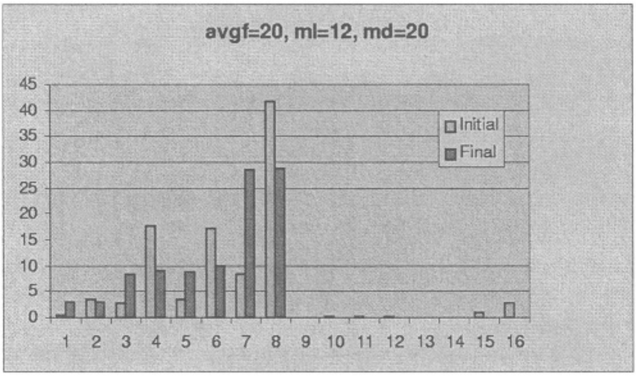


Fig. 5. Product market shares in the case of consumers talking to about 20 friends about their shopping (avgf=20), remembering at most 12 products (ml=12) and with memory duration of 20.

This can be interpreted that having many friends or collecting many opinions among the same market does actually empower the customer in making the best selection. It is interesting to note that the only initial advertisement cannot compensate for the further product comparisons communicated among the consumers. However, the initial product advertising effort results in the consumers remembering and, then, choosing the more advertised products among the low priced ones.

An alternative scenario would be to keep an average number of friends equal to 0, but increase the consumer memory limit to 12, Fig. 4.

In this case, the initial and final distribution look alike and tend to converge to the ideal market shares distribution but a bias toward the products investing in the initial advertising is evident.

Finally, if both the average number of friends (avgf=20) and the memory limit (ml=12) increase, then the initial and final distribution differ, the final one most closely matching the ideal ones, Fig. 5.

Comparing the initial and final distributions of market shares it appears that exchanging information about products with friends and remembering a number of them is the key to make a successful choice in this scenario. Indeed this observation is at the very base for the development of several strategies to deal with comparative on-line shopping.

4 Conclusion

An agent based methodology and tool to study market behaviors under several conditions of information diffusion has been described. The reported experimentation, in the context of an hypothetical economy, shows how this approach can be used to analyze and visualize market shares resulting after many complex information-based interactions among economic agents. Concerning electronic shopping and, especially, comparative shopping engines, the reported experiments show the significance of exchanging information among economic agents. Indeed, this is the key to make good/bad buying choice. Obviously, buyers and sellers regards each choice from a different perspective. This observation and this approach can help the development of novel marketing strategies in the comparative on-line shopping environment. As well as to further enrich the tools available for studying consumers' decision making.

References

1. Bettman, J. (1979). *An information processing theory of consumer choice.* Addison Wesley, Reading, MA (USA).
2. Bettman, J., Luce, M., and Payne, J. (1988). Constructive consumer choice processes. *Journal of Consumer Research,* pages 187–217.
3. Brian, A. W. (1994). Inductive reasoning and bounded rationality. *American Economic Review,* pages 406–411.
4. Currim, I. (1988). Disaggregate tree-structured modeling of consumer choice data. *Journal of Marketing Research,* pages 253–265.
5. Degeratu, A. M., Arvind, R., and Wu, J. (2000). Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes. *International Journal of Research in Marketing,* pages 55–78.
6. Guadagni, P. and Little, J. (1983). A logit model of brand choice calibrated on scanner data. *Marketing science,* pages 203–238.
7. Hales, D. (1998). An open mind is not an empty mind - experiments in the metanoosphere. *The Journal of Artificial Societies and Social Simulation (JASSS).*
8. Hoyer, W. (1988). An examination of consumer decision making for a common repeat purchase product. *Journal of Consumer Research,* pages 822–829.

9. J. G. Lynch, J. and Ariely, D. (2000). Wine online: search costs and competiotion on price, quality and distribution. *Marketing Science,* pages 1–39.
10. Kephart, J. O., Hanson, J. E., Levine, D. W., Grosof, B. N., Sairamesh, J., Segal, R., and White, S. R. (1998). Dynamics of an information-filtering economy. In *Cooperative Information Agents,* pages 160–171.
11. Lomuscio, A., Wooldridge, M., and Jennings, N. R. (2001). A classification scheme for negotiation in electronic commerce. In *AgentLink,* pages 19–33.
12. Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM,* pages 31–40.
13. Neri, F. (2001). An agent based approach to virtual market place simulation. In *Congresso dell'Associazione Italiana Intelligenza Artificiale 2001 (AIIA01),* pages 43–51.
14. Neri, F. and Saitta, L. (1996). Exploring the power of genetic search in learning symbolic classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* PAMI-18:1135–1142.
15. Rocha, A. P. and Oliveira, E. (1999). Agents advanced features for negotiation in electronic commerce and virtual organisations formation process. *Agent Mediated Electronic Commerce - An European Perspective,* LNAI 1991.
16. Rodriguez-Aguilar, J. A., Martin, F. J., Noriega, P., Garcia, P., and Sierra, C. (1998). Towards a test-bed for trading agents in electronic auction markets. *AI Communications,* 11(1):5–19.
17. Sierra, C., Jennings, N., Noriega, P., and Parsons, S. (1998). A framework for argumentation based negotiation. In *Intelligent Agents IV,* volume LNAI 1365, pages 177–192, Vienna, Austria. Springer-Verlag.
18. Smith, M. D., Bailey, J., and Brynjolfsson, E. (2001). Understanding digital markets: review and assessment. *Draft available at http:ecommerce.mit.edupapersude,* pages 1–34.
19. Viamonte, M. J. and Ramos, C. (1999). A model for an electronic market place. *Agent Mediated Electronic Commerce - An European Perspective,* LNAI 1991:3–28.
20. Wright, P. (1975). Consumer choice strategies: symplifying vs optimizing. *Journal of Marketing Research,* pages 60–67.

Metanarratives and Believable Behavior of Autonomous Agents

Mirko Petric[1], Inga Tomic-Koludrovic[2], and Ivica Mitrovic[3]

[1]Department of Visual Communication Design
Arts Academy, Split
mirko.petric@umas.hr
[2]Department of Sociology
University of Zadar
inga.tomic-koludrovic@umas.hr
[3]Department of Computer Science
Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Split
ivica.mitrovic@umas.hr
UMAS, University of Split
Glagoljaska bb
21000 Split, Croatia

Abstract. The aim of this conceptual paper is to propose separating social attitude engines from emotion engines of autonomous agents, as one way of increasing their social intelligence and consequently the believability of their behavior. The aim of the proposed separation is to clearly distinguish between the social and psychological aspects of agent behavior. In the existing emotion engines, the two aspects are blended to a degree which frequently prevents modelling of the elements of complex social interactions found in contemporary society. Our view is that the development of the proposed separate socio-political modules of social attitude engines could enable introduction of political and ideological elements into agent behavior. One way of introducing these elements into the agents' social attitude engines is via their narrative knowledge. In order to accomplish this, Jean-Francois Lyotard's notion of "metanarratives" has been used in this paper, as well as Fredric Jameson's reinterpretation of that notion. Three globally recognizable ideal types (neo-liberal, fundamentalist, and alternative) have been supplied with narratives translated into a conceptual model applicable in modelling of conversational agents. In the future, the presented socio-political attitude model should be expanded by means of addition of attitudes from various other areas of social life, in order to develop complex social attitude engines.

1 Introduction

The research this paper is based on was originally inspired by interactions of autonomous agents developed within the Carnegie-Mellon University's OZ project [3], [4], [5], [22], [27], [33] and Servant/Master scenarios developed within the Stanford University's Virtual Theater project [17]. More recently, the authors have discovered a particular affinity with the research program of the Socionics project[25] and the concept of embodied conversational agents [10]. Our focus is currently on

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 145–155, 2004.

providing theoretical foundations for developing believable behavior of socially intelligent agents.

In Dale's taxonomy [11], which represents an adaptation and extension of Wooldridge and Jennings [40], the agents that are at the center of our field of interest are described as the user interface agents and believable agents.

The aim of this conceptual paper is to propose separating social attitude engines from emotion engines of autonomous agents, in order to increase their social intelligence[1] and consequently the believability of their behavior.

Namely, in the field of modelling of believable agents, the elements that various researchers describe as personality traits, moods and attitudes [35], as well as emotions, features, attitudes, standards and behaviors [5], have thus far been blended in the agents' emotion engines[2].

Examples of layered (hybrid agent) architecture that model social competences separately (e.g. InteRRaP [29]) have been devised for systems whose focus is neither embodiment nor believability. These systems are goal-oriented and define social competences in a very limited manner.

In the field of modelling of believable agents, even where researchers distinguish between emotional and social aspects of the agents, they are modelled from the same substructure of agent architecture. The dominant modelling approach has been socio-psychological: what elements of the agents' social worlds there exist are approached from a psychological perspective. This is a logical outcome of the initial modelling strategies based firmly on modular cognitive psychological approaches.[3]

A recent approach [37], stimulating in the field of modelling of believable agents, has elaborated a PECS model in which the variables of the physical, emotional, cognitive and social states are presented separately. The social aspect of this model is based on the notion of "social position", as are the role concepts for agents developed by Lindemann-von Trzebiatowski and Münch [21], Guye-Vuillème and Thalmann [16] or Prendinger and Ishizuka [32]. However, if the primary goal of modelling activity is increasing the agents' believability, the concepts inflexibly connecting social positions and social roles are at a disadvantage of not being associated with the contemporary social trends.

In this regard, it is worth noting that social stratification based on social position, as well as fixed and consistent relationships between social positions and social roles, are characteristic of traditional and modern societies. Contemporary society is much

[1] We use the term "social intelligence" following Dautenhahn [12], [13l and by analogy with the use of the terms such as "narrative intelligence" [7]. To avoid "the somewhat difficult [task] to define [the] concept of social intelligence", other researchers have used the term "social competence" [36]. While we recognize that the term is difficult to define, we use it to immediately situate the research into the field of artificial intelligence.

[2] We have adopted the expression "emotion engine" after Elliott's [15] and similar uses of the term. Rousseau and Hayes-Roth [35] define the psychological and social aspects of agents through a "social-psychological model", and Bates et al [5] speak of an "EM architecture" defining emotion and social relationships.

[3] The emotional models of autonomous agents are most frequently based on the cognitive theory of emotion of Ortony, Clore and Collins [30] and Elliott's [14] implementation of OCC model in multiagent systems.

more differentiated and pluralistic, and organized in a more flexible way. In a postmodern context, personal identity is not firmly tied anymore to neither the social position nor the social role.

The society we refer to here as "postmodern" has been described by Beck [6] as "risk society", coming about as a consequence of the process of "reflexive modernization". According to Beck, when modernization reaches a certain level it entails a changing relationship between socials structures and social agents, in which agents become decreasingly constrained by structures.

Agents become increasingly individualized and have to think through traditional social roles by themselves (hence the expression "reflexive modernization"). In contemporary societies, agents create (and are even structurally forced to create) their own biographies and do not simply follow traditional, socially prescribed roles. In such a context, traditional roles such as "servant" and "master" (used in believable agent modelling, partly due to present-day technical limitations) tend to appear antiquated in terms of their present-day social relevance. We feel that believable agent modelling should also take into account the recent processes in social development.

In a postmodern context, identities are constructed and modified in the course of communication practices. When modelling, a way has to be found to increase the agents' believability in such a social context.

This paper proposes that this should be done by means of separating the agents' social attitude engines (containing social beliefs and values) from their emotion engines, in order to increase the social complexity of the agents' behavior. The two engines are conceptualized as connected in some respects (e.g. in their simultaneous response to a stimulus) but also - by virtue of their separate conceptualization - as enabling different responses to similar stimuli in different agents.

The development of separate emotion and social attitude engines would also make it possible to introduce political and ideological elements into agent behavior. These elements - however elementary in practical execution at the present-day level of technical development - would secure further differentiation of the agents' social worlds.

One way of introducing these elements into the agents' social attitude engines is via their narrative knowledge. As suggested by Petric [31], Jean-Francois Lyotard's notion of "metanarratives"[4] has been used in this paper, because it enables a smooth connection of individual narrative histories and wider interpretations of the nature of the social bond.

[4] The original Lyotard's term "grand récit" [23] has been translated as "grand narrative" in the American edition of his book [24]. The term "metanarrative" can be used interchangeably with this term, as was done by the author himself. Jameson's term "master narrative" [24] is an apt description of the implications of both these terms.

2 Separating Social Attitude Engines from Emotion Engines

Researchers in the field of believable agent modelling, as it developed in the 1990s, based their work on various socio-psychological approaches. We suggest in this paper that the social and pyschological aspects of agent architecture could be separated, in an attempt to model socially more complex agents. In the model that we propose, the separation of the agents' social attitude engines and emotion engines would increase modularity and contribute to believability defined in social terms.

In the model proposed here, the psychological aspects of the agent are defined from the agent's emotion engine, while the agent's social aspects are defined from a separate social attitude engine. We hold it that the agent's behavior is a consequence of not only psychological but also of socially defined parameters, and that the latter can be represented by means of a reduced definition of attitude.

In sociological research, attitudes are "[v]ariously defined as an orientation towards a person, situation, institution, or social process, that is held to be indicative of an underlying value or belief; or, [...] as a tendency to act in a certain (more or less consistent) way towards persons and situations" [26]. In the field of believable agent modelling, the sociological research of attitudes is especially stimulating because, in this research, "[a]ttitudes [...] are (sometimes) assumed to predict behaviour" [26].

In simplest terms, an attitude is a positive or negative evaluation of someone or something, which causes a certain type of behavior. "At the simplest level, attitude questions invite people to agree or disagree, approve or disapprove, say Yes or No to something"[5] [26]. In the proposed model, the agent's social attitude engine is based on attitudes reduced to a set of simple oppositions ("+", "-"), and an absence of attitude ("0").

This social attitude engine is conceived of as separate from but related to an emotion engine, such as, for instance, an emotion engine based on Elliott's interpretation of the OCC model [14]. In agent interactions, positive, negative, and neutral evaluations in the social attitude engine trigger off appropriate responses from the agent's emotion engine. This makes it possible to keep clear "a distinction [...] between attitudes and the emotions that they generate" as advised by Reilly and Bates [33]. However, in the proposed model, attitudes do not "represent personal tastes and preferences", as is the case in Reilly's and Bates's description. Attitudes are conceived of here as representative of the wider social structure within which they are positioned and which determines them[6]. In other words, the approach to the agent's attitudes in the proposed model departs from and emphasizes their social believability rather than their psychological believability: agents and agent reactions represent social structures rather than individual entities.

[5] More sophisticated and well-established techniques for measuring attitudes include the Likert scale, the Thurstone scale, Osgood's semantic differential scale, the Bogardus social distance scale, and Guttman scales.

[6] Following Bourdieu [8], various tastes and preferences are not individually but socially structured. We intend to use this aspect of his model in the future development of the social attitude engines.

3 Attitudes and Narrative Knowledge: Metanarratives vs. Individual Narratives

In addition to the need for developing social intelligence, recent approaches to agent modelling also emphasize narrative intelligence [12], [28]. Namely, stories are considered to be “fundamental to human (social) intelligence” [12] and humans are seen as “autobiographic agents and life-long learners [...] constantly re-telling and re-interpreting their autobiography and their interpretation of the world” [12].

By telling stories, people make sense of the world, they “order its events and find meaning in them by assimilating them to more-or-less familiar narratives” [28]. Needless to say, at the present level of development, AI research into what Blair and Meyer [7] label as “narrative intelligence” and define as “human ability to organize experience into narrative form” takes place at a rather elementary level (when compared with the complexity of human narrative knowledge) and concerns primarily the agent’s individual experience.

However, this does not mean that human narrative intelligence cannot be mirrored in agent modelling. Mateas and Sengers [28] recount a proto-story in which one of the authors of the paper was involved when she and her friend were “barely verbal”. “Phoebe! Pizza! Phoebe! Pizza!” was a story told by Sengers’s two-year-old friend when she happened to arrive simultaneously with the pizza delivery boy. According to the authors, this story meant approximately, “Can you believe it? Phoebe and pizza came into the house at the same time!”

In a sociologically-minded approach, the attitudes that an agent holds - since they are always indicative of an underlying value or belief - can be conceived as a link between the individual experience and a wider social structure. What’s more, as sociology of culture tells us, even those attitudes and behaviors that appear highly individual are in fact predisposed by “the knowledge gained from living in a particular culture” [20].

In the field of modelling narratively intelligent and socially believable agents, the problem would, then, be in relating the individual experience to the model of knowledge imposed by the social context. In our conceptual model, this latter model of knowledge is placed in the agent’s social attitude engine.

In contrast with the agent’s individual narrative knowledge, the knowledge that is related to the interpretation of the nature of the social bond could be described by means of what Lyotard [23] refers to as grand narratives or metanarratives. These narratives include the agent’s ideological and political attitudes and are therefore obviously aimed at explaining social issues at a supraindividual level.

Although Lyotard [23] actually discussed the problems of the legitimation of knowledge and scientific research, his insights on the nature of the two great legitimizing narrative archetypes of modernity (“that of the liberation of humanity and that of the speculative unity of all knowledge”) [18] can also be viewed as related to wider social issues. As Jameson [18], puts it: “Doing science”[...] involves its own kind of legitimation [...] and may therefore be investigated as a subset of the vaster political problem of the legitimation of a whole social order” [18].

Paraphrasing Lyotard, we could say that every social attitude is “obliged to legitimate the rules of its own game” [24]. Furthermore, as noted by Jameson [18], the elements of the legitimizing metanarratives are nowadays buried in the “political unconscious” of every individual “as a way of ‘thinking about’ and acting in our current situation”. Since these accounts of our political beliefs are essentially products of what originally used to be philosophical discourse, they are highly consistent and therefore stimulating in the context of agent modelling.

4 From Narratives to Models: An Outline of Socio-political Ideal Types

In order to create a conceptual model on the basis of the narratives as described above, one needs to translate the attitudes that are still too complex in the agent modelling context into a more applicable form. One way of doing this is by utilizing the sociological notion of the ideal type, introduced by Weber [39], in order to enable the sociologists to see the social world in a more systematic way.

It goes without saying that “societies differ in many ways from their respective types” [19], but the implied oversimplification of reality “in order to bring out certain of its most important features” [19] is simply necessary in the context of agent modelling, even more so than in the original context of sociological research.

In order to achieve believability of socio-political attitudes in a contemporary context, we have constructed three globally recognizable ideal types and supplied each of them with a narrative that can be translated into a conceptual model applicable in agent modelling. We have labeled these types as neo-liberal, fundamentalist, and alternative, and provided them with the following metanarratives:

Neo-liberal metanarrative: “The market is the most important thing in the world. Everything is and should be organized according to the market principle. Market equals freedom. The market is just in itself: redistribution of wealth is not necessary. Everybody should be an entrepreneur and compete on the market. Individuals should be mobile and flexible, and disregard the needs of the collective. The nation is not more than a big corporation. There should not be any obstacles to free trade: the market is and should be global. Military intervention is justifiable in cases where freedom is endangered. The furtherance of material progress is important. Family, nation, one’s religion, natural resources are of secondary importance to commercial values. One should not think about the long-term consequences of one’s actions: the present moment is everything.”[7]

Fundamentalist metanarrative: “Market and materialism are evil. The needs of the individual are secondary to the needs of the collective. Everything should be done and all the resources of the community should be used to preserve the inherited collective

[7] The neo-liberal metanarrative, as presented here, is based on Bourdieu [9].

values. One should fight against everything that puts them in danger: commercial values, materialism, selfish individualism. It is important to preserve the institutions of the past: one's family, nation, religion. The past provides solutions to the problems. One should take sides in a cosmic war between the forces of good and bad. The instructions of charismatic leaders and sacred texts should be taken literally and rigidly followed."[8]

Alternative metanarrative: "Individuals should be mobile and flexible, but also heed the needs of the collective. The market is not just in itself: social solidarity and redistribution of wealth should amend it. Commercial values and materialism are not the most important thing. Natural resources should not be destroyed because of material progress. One should think about the future more than about the past. Economies should be local. Individuals should behave tolerantly and not resort to violence. One should respect other people's beliefs, lifestyles, sexual orientation and religion."[9]

Like all the sociological ideal types, these labels are obviously an oversimplification of reality. They invite criticism even more so than has previously been the case, because they have been abstracted from a more complex, highly pluralistic and contextually dependent ("postmodern") social reality. On the other hand, we hold it that the three proposed types are globally recognizable, and have therefore opted for them rather than for essentially simpler and more manageable sets of socio-political attitudes recognizable and believable in a much narrower socio-political context (e.g. attitudes of the constituencies of U.S. Republican and Democratic parties).

It should be noted that, because of the wish to achieve a wider scope of reference and flexibility, the term "fundamentalist" should not be taken to denote merely "Islamic fundamentalist". It can find application in that particular context, but can also extend beyond it. Similarly, the label "alternative" has been chosen rather than e.g. the term "anti-globalist", which is both narrower in reference and more ambiguous ("anti-globalism" can be both left-wing and right-wing, while "alternative" is unequivocally left-wing). Another complication with the use of the term "anti-globalist" would be that, in contrast with the label "alternative", left-wing anti-globalism does not only imply an individual position but also a collective actor (movement) which includes the possibility of violent resistance.

[8] The fundamentalist metanarrative is largely based on Armstrong [2].

[9] The alternative metanarrative is based on Tomic-Koludrovic [38]. It is essentially a summary of attitudes characteristic of "alternative" " approaches to the social world in the second half of the 1980s and in the 1990s. The recent anti-globalist movement is currently given more media coverage, but various "alternative" approaches continue to exist and are globally recognizable.

The following statements on a set of issues connected with the outlined metanarratives are presented here in the form of a table:

Issue	Neo-liberal	Alternative	Fundamentalist
Market	Market is the most important thing in the world.	Market should be regulated in order to ensure social justice.	Market is evil.
Individualism	Individual ability and individual needs come first.	It is important both to be individual and not to be selfish.	The needs of the individual are secondary to those of the collective.
Flexibility	One should be limitlessly flexible.	Flexibility is important within certain limits.	Inherited values should be rigidly followed.
Tradition	The present moment is everything.	One should think about the future more than about the past.	The past provides solutions to the problems.
War	War is justifiable when freedom is endangered.	War is never justifiable.	One should take sides in a cosmic war between good and bad.
Religious tolerance	Religion is not a problem unless it interferes with the market.	One should respect other people's beliefs and religion.	Adherence to the principles of one's religion is the only right thing to do.
Materialism	The furtherance of material progress is important.	Materialism should be restricted because it contributes to the waste of natural resources.	Materialism is evil.
Redistribution of wealth	There should be no social solidarity.	Redistribution of wealth is important because it ensures a minimum of social justice.	All the resources of the community should be used to preserve the values of the collective.

These statements are indicative of the agents' attitudes toward various issues, and can be presented in the following table, in which the plus sign indicates a positive attitude, a minus sign a negative attitude, and zero sign a neutral attitude.

Issue	Neo-liberal	Alternative	Fundamentalist
Market	+	0	-
Individualism	+	0	-
Flexibility	+	0	-
Tradition	-	0	+
War	0	-	+
Religious tolerance	0	+	-
Materialism	+	0	-
Redistribution of wealth	-	+	0

In the course of further elaboration, a set of statements with more detailed expressions of the basic attitude will be added to every cell in the table (i.e. to every plus, minus or zero sign). These sets of statements will then be used as a basis for the development of the socio-political attitude modules of the social attitude engines of believable conversational agents.

5 Concluding Remarks

The aim of the separation of social attitude engines and emotion engines of autonomous agents proposed in this paper has been to clearly distinguish between the social and psychological aspects of agent behavior. Furthermore, within the area of the social, we have concentrated on attitudes, defined as positive, negative or neutral orientations towards a social process and taken to be indicative of an underlying belief.

Socio-political attitude models resulting from this separation and based on the elements of metanarratives legitimating the nature of the social bond are meant to serve as a basis for a future development of social attitude engines for conversational agents. The agent behavior resulting from these engines is conceptualized as neither determined by a role nor goal-driven.[10]

Our model is not role-oriented because our aim is to increase believability in the contemporary social context, in which fixed roles are decreasingly important as a basis for social identity of actors. Likewise, the aim of the conversational agents that could be developed from our model would not be goal-driven, but would be to present their different socio-political attitudes (i.e. their socio-political identities) in a conversation.

To apply goal-driven agent concept leading to a change of attitude would be extremely difficult in this particular context, since the presented socio-political attitudes consist of a large number of highly consistent and abstract elements, and their change would depend on a number of complex social factors and experiences difficult to simulate in agent environments.

However, conversational interactions - taking place exclusively among artificial autonomous agents or involving an active human user as well - could help the interested users learn about the outlined socio-political metanarratives and open up new lines of thought in the form of creative play.

In the context of agent modelling, the proposed separation of social attitude engines and emotion engines could hopefully result in greater social believability. In order to achieve this, the presented socio-political attitude model should be expanded by means of addition of attitudes from various other areas of social life.

Obviously, there is still a fair amount of work to be carried out to bridge the gap between the insights of the background disciplines referring to human societies and

[10] The role and goal concepts of agent behavior are frequently closely connected. For instance, in the Inhabited Market Place scenario [1] agents in the roles of the seller and buyer lead goal-oriented conversations. There are, however, also models which are purely goal-driven (e.g. [34]).

present-day technical possibilities in the field of believable agent modelling. However, we are of the opinion that the outline of the conceptual model presented in this paper should not be relegated by the modelling community to the status of an intra-sociological debate. If believability of socially intelligent agents is to be achieved in complex present-day social circumstances and if the agents are to assume a part of the richness and complexity these circumstances give rise to, even if only in extremely limited interactions, the modelling of these interactions should rest on the solid theoretical foundations of the disciplines trying to account for contemporary societal trends.

References

1. André, E., Rist, T., Mulken, S. v., Klesen, M., Baldes, S.: The Automated Design of Believable Dialogues for Animated Presentation Teams. In: Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (eds.): Embodied Conversational Agents. The MIT Press, Cambridge MA and London (2000) 220–255
2. Armstrong, K.: The Battle for God: Fundamentalism in Judaism, Christianity and Islam. Harper-Collins, London (2000)
3. Bates, J.: Virtual Reality, Art, and Entertainment. Presence: The Journal of Teleoperators and Virtual Environments 1 (1992) 133–138
4. Bates, J.: The Role of Emotion in Believable Agents. Technical Report CMU-CS-94-136. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (1994)
5. Bates, J., Loyall, A. B., Reilly, W. S.: An Architecture for Action, Emotion, and Social Behavior. Technical Report CMU-CS-92-144, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (1992)
6. Beck, U.: Risikogesellchaft: Auf dem Weg in eine andere Moderne. Suhrkamp, Frankfurt/M(1986)
7. Blair, D., Meyer, T.: Tools for an Interactive Virtual Cinema. In: Trappl, R., Petta, P. (eds.): Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents. Springer-Verlag, Berlin (1997)
8. Bourdieu, P.: La distinction: critique sociale du jugement. Minuit, Paris (1979)
9. Bourdieu, P.: The Essence of Neo-Liberalism. Le Monde Diplomatique, <http://mondediplo.com/1998/12/08bourdieu>(1998)
10. Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (eds.): Embodied Conversational Agents, The MIT Press, Cambridge MA London (2000)
11. Dale, J.: A Mobile Agent Architecture for Distributed Information Management. Ph.D. thesis, University of Southampton (1997)
12. Dautenhahn, K.: The Art of Designing Socially Intelligent Agents - Science, Fiction and the Human in the Loop. Applied Artificial Intelligence, Special Issue on Socially Intelligent Agents (1998) 7–8
13. Dautenhahn, K., Bond A., Cañamero, L., Edmonds, B.: Socially Intelligent Agents: Creating Relationships with Computers and Robots. In: Dautenhahn, K., Bond, A. H., Cañamero, L., Edmonds, B.: Socially Intelligent Agents: Creating Relationships with Computers and Robots. Kluwer Academic Publishers (2002)
14. Elliott, C.D.: The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System. Ph.D. diss. Northwestern University, Evanston, Illinois (1992)
15. Elliott, C.D.: Research problems in the use of a shallow Artificial Intelligence model of personality and emotion. In: Proceedings of the Twelfth National Conference on Artificial Intelligence. AAAI, Seattle, WA (1994) 9–15
16. Guye-Vuillème, A., Thalmann, D.: A High-Level Architecture for Believable Social Agents, VR Journal, 5 (2001) 95–106

17. Hayes-Roth, B., Van Gent, R., Huber, D.: Acting in Character. In: Proceedings of the AAAI Workshop on AI and Entertainment (1996)
18. Jameson, F.: Foreword. In: Lyotard, J.-F.: The Postmodern Condition: A Report on Knowledge. University of Minnesota Press, Minneapolis and London (1984)
19. Johnson, A. G.: The Blackwell Dictionary of Sociology. Blackwell, Oxford and Molden, MA (2000)
20. Kirby, M., Kidd, W., Koubel, F., Barter, J., Hope, T., Kirton, A., Madry, N., Manning, P., Triggs, K.: Sociology in Perspective. Heinemann, Oxford etc (2000)
21. Lindemann-von Trzebiatowski, G., Münch, I.: The Role Concept for Agents in Multi-Agent Systems. Sozionik aktuell 3 (2001) 15–30
22. Loyall, A. B., Bates, J.: Personality-Rich Believable Agents That Use Language. In: Proceedings of the First International Conference on Autonomous Agents, Marina del Rey, California (1997)
23. Lyotard, J.-F.: La condition postmoderne. Minuit, Paris (1979)
24. Lyotard, J.-F.: The Postmodern Condition: A Report on Knowledge, University of Minnesota Press, Minneapolis and London (1984)
25. Malsch, T.: Naming the Unnamable: Socionics or the Sociological Turn of/to Distributed Artificial Intelligence. Autonomous Agents and Multi-Agent Systems 4 (2001) 155–186
26. Marshall, G.: Oxford Concise Dictionary of Sociology. Oxford University Press, Oxford and New York (1994)
27. Mateas, M.: An Oz-Centric Review of Interactive Drama and Believable Agents. Technical Report CMU-CS-97-156. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA (1997)
28. Mateas, M., Sengers, P.: Narrative Intelligence. In: Proc. Narrative Intelligence, AAAI Fall Symposium 1999, AAAI Press, Technical Report FS-99-01 (1999)
29. Muller, J. P., Pischel M.: The Agent Architecture InteRRaP: Concept and Application. Technical Report RR-93-26, DFKI Saarbrucken (1993)
30. Ortony, A., Clore, G. L., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, Cambridge (1988)
31. Petric, M.: Missing Narratives: The Notion of "Grand Récit" in Artificial Agent Modelling. Paper delivered at Society of Literature and Science Annual Conference, Buffalo, NY (2001)
32. Prendinger, H., Ishizuka, M.: Social Role Awareness in Animated Agents. In: Proceedings 5th International Conference on Autonomous Agents (Agents-01) (2001)
33. Reilly, W. S., Bates, J.: Building Emotional Agents. Technical Report CMU-CS-92-143, School of Computer Science, Carnegie Mellon University (1992)
34. Rizzo, P., Veloso, M., Miceli, M., Cesta, A.: Personality-Driven Social Behaviors in Believable Agents. In: Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents. Cambridge, Massachusetts (1997)
35. Rousseau D., Hayes-Roth, B.: A Social-Psychological Model for Synthetic Actors. In: Proceedings 2nd International Conference on Autonomous Agents (Agents'98) (1998)
36. Schillo, M., Allen, S., Fischer, K., Klein, C. T.: Socially Competent Business Agents with Social Competence: Using Habitus-Field Theory to Design Agents with Social Competence. In: Proceedings of the AISB'00 Symposium on Starting from Society - the Application of Social Analogies to Computational Systems, AISB, Birmingham, UK (2000)
37. Schmidt, B.: Agents in the Social Sciences - Modeling of Human Behavior. Sozionik aktuell 3 (2001) 5–14
38. Tomic-Koludrovic, I.: Alternativna kultura kao oblik otpora u samoupravnom socijalizmu. Drustvena istrazivanja 4–5 (1993) 835–862
39. Weber, M.: Wirtschaft und Gesellschaft. Mohr, Tübingen (1921)
40. Wooldridge, M. and Jennings, N. R.: Intelligent Agents: Theory and Practice. In: Knowledge Engineering Review, 10 (1995) 115–152

FORM – A Sociologically Founded Framework for Designing Self-Organization of Multiagent Systems*

Michael Schillo[1], Klaus Fischer[1], Bettina Fley[2], Michael Florian[2],
Frank Hillebrandt[2], and Daniela Spresny[2]

[1] German Research Center for Artificial Intelligence (DFKI),
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
{schillo, kuf}@dfki.de
[2] Department of Technology Assessment, Technical University of Hamburg-Harburg,
Schwarzenbergstr. 95, 21071 Hamburg, Germany
{bettina.fley, florian, hillebrandt,daniela.hinck}@tu-harburg.de

Abstract. We propose *FORM,* a new *Framework for self-Organization and Robustness in Multiagent systems.* This framework supports the design of task-assignment multiagent systems in a way that is informed by sociological theory. It is founded on the habitus-field-theory of sociologist Pierre Bourdieu. In accordance to this theory, we consider the special quality of "organization" as an autonomous and self-organizing social entity with clear distinction to the coordination via social interactions. Organizations are viewed as both "autonomous social fields" and "corporate agents", which are competing with other organizations in the same domain. While our framework makes no claims about an underlying agent architecture, it consists of a matrix of mechanisms for delegation (task delegation and social delegation), which we consider a central concept to define organizational relationships. Using this matrix as a basic toolset, we propose a spectrum of seven types for the structure of multiagent systems, defined by qualitatively different relationships.

1 Introduction

Since the 1980's, multiagent systems (MAS) have been an influential research strand in the artificial intelligence community. In the pursuit of a new paradigm, researchers worked on MAS, which promise to be "more *reliable* than are centralised systems" ([1]: 9, citing [2]), to have significant advantages, e.g. "more flexibility ... and increased reliability " ([3]: 5), and finally, to offer useful features such as "parallelism, robustness and scalability" ([4]: 1). In a highly influential article, Jennings writes that "the development of robust and scalable software systems requires autonomous agents that can complete their objectives while situated in a dynamic and uncertain environment, that can engage in rich, highlevel social interactions, and that can operate within flexible organisational structures" [5]. The advantages of agents that act in organisational structures he sees are that organizations can encapsulate complexity of subsystems (simplifying representation

* This research is supported by the Deutsche Forschungsgemeinschaft in the *Socionics* framework under contracts Si 372/9-2 and Fl 336/1-2.

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 156–175, 2004.

and design) and modularize functionality (providing the basis for rapid development and incremental deployment). These aspects are captured by *holonic MAS* as proposed by Fischer et al. [6].

There are several more reasons to believe that enabling agents to such a kind of behaviour will have a positive effect on a MAS. One is that organizations are ascribed the ability to overcome the limitations of individuals and coordinate their actions in such a way that the organization as a whole achieves higher performance. Another is that an organization can be more persistent than a group of interacting agents because of formal structures that regulate membership, procedures, aims of the group, and other constraints which are important to organize joint action. This is achieved e.g. by the separation of ends from motivations for paid members in organizations where money acts as motivation. Also, in contrast to systems based merely on ad hoc interaction, organizations do not fall apart as soon as an agent stops to interact. Finally, organizations institutionalize anticipated co-ordination, which can lead to the reduction of communicational effort.

Turner and Jennings [7] choose the notion of organization to work on scalability issues in MAS, a topic where this notion plays an important role. They improve system performance by the individual agents' ability to determine the most appropriate communication structure for the system at run-time by themselves and to change this structure as their environment changes. The work of So and Durfee [8] is similar but restricts analysis of tree-like structures to the performance in homogeneous MAS. In their work, all communication links between the agents are of the same nature and organization focuses on the arrangement of communication channels rather than (re-)defining the nature of each channel. Other work reports on agents creating, joining or leaving firms depending on respective utility [9], or several agents that merge into a single agent according to the tasks to be performed [10]. Note that there are other links between DAI and the term "organization", we have not mentioned yet. One link is the use of agents for modeling human organizations (cf. e.g. [11]), others use MAS for business process engineering (e.g. [12]), a third link is the process of structuring reactive MAS inspired by biology (e.g. pheromon-based computation). However, in contrast to these uses of the term, we prefer to use organization in a sociological sense as inspiration for the creation of problem solving MAS.

Looking at this literature shows a concentration on selected organizational aspects like communication topologies and the transformation of the members of the organization into a single agent. The organizational structure here is solely used to determine and canalize communication and interactions. An interesting commonality is the use of the term organization to define self-organization. In this sense, self-organization means the process of generating social structure, which is the result of individual choices by a set of agents to engage in interaction in certain organizational patterns, depending on their own resources and the environmental context. A definition we are ready to adopt for the remainder of this discussion.

From a sociological point of view, we agree with the suggestion that self-organization means the process of generating, adapting and changing organizational structure, but we do not agree with the definitions of the term organization in the discussed area of DAI discussed so far because of two reasons: Firstly, in a sociological sense, organizations are specific social entities. They differ from other social phenomena (like interactions, group behaviour, etc.) as their social structures are formalized to a great extent. Formal structures regulate constraints (i.e. rules of organizational membership, goals, operational procedures, social norms, etc.) to control the agents' behavior. Secondly, organizations are more than the sum of interactions between single agents. Organizations are emergent phenomena. Their social structure evolves, exists, and persists independently from single agents' intentions and goals, even though they arise from the actions of individuals. In addition, these structures react upon the goals and actions of agents at the same time. Thus, organizational structures may be considered as a result of individual choices by a set of agents to engage in interaction in certain organizational patterns. Nevertheless, the agents' behaviour is constrained by formal structures as soon as agents become members of an organization. For the scope of this paper, the term self-organization will refer to this sociological meaning of organization. In general, organization in a sociological sense is only one example for self-organizing processes by the interactions of single agents. However, we suggest that organizations are a particularly interesting subject for DAI because of the tension between the social control due to formal structures and the deliberate behavior of self-interested agents becoming member of an organization.

The main objective of this paper is to use a (in a sociological sense) more advanced model of organization and self-organization to create a framework for MAS. We are particularly interested in one important type of MAS, namely task-assignment MAS [13], i.e. we assume agents to engage in interaction with other agents in order to distribute tasks according to costs, competence, maybe even task load. The MAS as a whole is supposed to solve the problem of assigning tasks to agents according to these measures.

Following this introduction, we lay out the foundations for our work. Section 2.1 reviews the concept of *holonic MAS* which has proven to be useful as an abstract framework to describe organization in MAS. Section 2.2 deals with the concept of organization based on the habitus-field-theory of Pierre Bourdieu. We then describe in Section 3 the *Framework for self-Organization and Robustness in Multiagent systems (FORM),* which combines the concept of holonic MAS and the concept mentioned of organization to describe self-organization for task-assignment MAS.

2 A Sociological Approach to Self-Organization in MAS: From Agent Interaction to Agent Organization

The problem of self-organization has been subject of numerous discussions concerning the question of the relationship between a system and its environment in various disciplines apart from DAI (cybernetics, biology, sociology, etc.). During the last decades, self-organization has become an established interdisciplinary notion (cf. e.g. [14]). The different theoretical approaches have in common that they call any kind of system "self-organizing", if it is able to preserve "operational closure" (i.e. autonomy) toward its environment and recursively determines its internal structures by itself as the environment changes (cf. Varela in [14]).

With reference to DAI, self-organization means that a MAS as a whole should be able to change its internal social structures (i.e. the creation of different organizational types) independently from designers' direct interventions and even independently from the intentions of self-interested single agents. From our point of view, self-organizing

MAS are confronted with a twofold challenge regarding their "autonomy" towards their environment: They have to deal with autonomy against external control of human designers as well as towards internal deliberation of self-interested agents. Firstly, they should be able to change their internal structures autonomously (i.e. by generating different organizational types for agent coordination and cooperation) independently from designers' interventions (external autonomy). Secondly, MAS should also be independent from harmful intentions and motivations of single agents (internal autonomy), although self-interested agents are the driving force enabling the self-organization in MAS. In this section we will firstly introduce the concept of holonic MAS in order to be able to describe the notion of organization in DAI terms. Secondly, we discuss our concept of organization, which is based on the sociological habitus-field-theory of Pierre Bourdieu.

2.1 Holonic MAS and Organization

In this work we restrict ourselves to MAS that are designed for task-assignment (cf. task-oriented domains [13]). Agents act in their environment in analogy to a market. The market consists of two sets of agents: providers and customers. Providers are agents that can perform tasks either through their capabilities or, alternatively, due to resources they have access to (database access, production resources for manufacturing domains, etc.). Tasks are of a certain type, have a deadline (latest delivery time), and may be composed of independent subtasks. Customers have tasks that should be performed, possibly they represent human users as avatars. We will not go into detail about what kinds of tasks are to be performed by the agents but rather concentrate on the effect of using a theory of organization for multiagent systems to achieve task-assignment. As long as provider agents are able to supply the tasks customers demand, they neither need to delegate tasks to other providers nor to cooperate. This may change if customers demand more complex services, which single agents cannot perform alone. This might be because the resources of any single agent are not sufficient, or the demanded services are compound of different kinds of tasks and no single agent is capable to perform all of them. In this case, provider agents need to carry out more complex and compound tasks jointly (delegate tasks to others).

To model these joint activities, the concept *holonic agent* or *holon* as defined by Fischer [15] is used. The concept is inspired by the idea of recursive or self-similar structures in biological systems [16]. A holonic superagent consists of parts called *body agents,* which in turn may be holonic agents themselves. Any holonic agent that is part of a whole, contributes to achieve the goals of this superior whole. The holonic agent may have capabilities that emerge from the composition of body agents and it may have actions at its disposal that none of its body agents could perform alone. The body agents can give up parts of their autonomy to the holon. To the outside, a holon is represented by a distinguished *head (agent)* which moderates the activities of the body agents and represents the holon to the outside. The advantages of this concept are threefold. Firstly, this technology preserves compatibitlity to MAS by addressing every holon as an agent, whether this agent represents a set of agents or not, is encapsulated. Secondly, as every agent may or may not represent a larger holon, holonic MAS are a way of introducing recursion to the modelling of MAS, which has proven to be a powerful mechanism in software design. Thirdly, the concept does not restrict us to a specified type of association between the agents, so it leaves room to introduce organizational concepts at this point.

As proposed by Fischer et al. [6], three types of association are possible for a holon: firstly, body agents can build a loose federation sharing a common goal for some time before separating to regulate their own objectives. Secondly, body agents can give up their autonomy and merge into a new agent. Thirdly, any nuance on the spectrum between the first and second scenario is possible, considering that agents can give up autonomy on certain aspects, while retaining it for others. In this case of flexible holons, the responsibility for certain tasks and the degree of autonomy that is given up is subject to negotiation between the agents participating in the holon, not a matter of pre-definition by the designer. However, what exactly the "nuances" or stages on this spectrum can be, has not yet been addressed and is the focus of this work. Throughout the remainder of this text, we will use the concept of holonic multiagent systems as the basic framework to specify organization in MAS.

2.2 Organizations as Autonomous Fields and Corporate Agents

In Section 1 we emphasized that the creation of organizations by the interactions of single agents can be considered as one very important example for self-organizing processes. Using the basic terms of Bourdieu's habitus-field theory, we will outline in the following a new conception of organizations as autonomous fields and corporate agents for an advanced model of self-organization in DAI. Our specific contribution to a sociological founded DAI is not only to define organizations as corporate agents, but also to consider them as *social fields* according to Bourdieu?s comprehension of the term. A fundamental characteristic of organizations that distinguishes them from any other kind of social field (macro-social fields, e.g. the economic field, micro-social fields, e.g. a group) is that they are formally organized or structured. These formal structures (programs, statutes, written rules, etc.) regulate aims of the organization, membership, division of labor, competence of members, distribution of profits, etc., in short: the task structure, the authority structure, the resource structure, etc. Nevertheless, it would be a contradiction to consider organizations as social fields, but reduce them only to their formal structures. According to Bourdieu, organizations are not static and formal apparatuses oriented towards a common function in which members fully adopt the aim of the organization mechanically as their own goal [17].

The term field within the theory of Bourdieu is an analytical category to describe the structural conditions for sociality (the social practice of agents) in general and in the matter of self-organization. Fields are attributed four characteristics which are important for the process of self-organization: Firstly, any field shows an objective structure of the relations between the social positions occupied by the agents acting in the context of a specific field. A position is defined by restrictions and possibilities it imposes upon agents, by the present and potential composition of all sorts of resources an agent possesses (in terms of Bourdieu: economic, cultural, social and symbolic capital), and by its relation to other positions. The agents need specific forms of cultural, economic, social and symbolic capital to take a specific position related to other positions in the field. Secondly, as any field can be compared to a game it follows its own "rules". These are, in contrast to a game like it is defined in game-theory, neither explicit norms to be obeyed by individuals nor the product of an intentional act, but regularities of practice. Thirdly, any field has its own logic, what makes it autonomous in comparison to other fields. For example, the interest of the economic field can be called “business is business” (i.e. making profits). This logic excludes practices which are proceeding in another logic [18], e.g. practices in politics that focuses on obtaining power. Last but not least, any field is a field of struggles. Bourdieu assumes that agents act in a field like players in a game. Like in a game, agents are opposing one another, they are interested in improving their relative positions in the field. In this sense, they are self-interested but in a specific way. The agent’s rationality depends on the forms of capital it possesses and must be defined as a practical sense for the game of the field (termed “illusio” by Bourdieu). Thus, their interests are socially shaped. As agents try to improve their relative positions the distribution of all species of capital, the regularities, and even the task structure of a social field can be object of the agents’ attempts to influence the structure of a field in favor of their socially structured interests. Therefore, we view the agent as the force behind the development, change and reproduction of social structure of any field.

Following Bourdieu, this means that the agents of an organization are interested in improving their relative positions in the organization. They act in a self-interested way. This might appear as an argument against using Bourdieu’s theory for modeling MAS, because in DAI-literature it is seen as an advantage that formal structures constrain self-interested agents to prevent opportunism. Nevertheless, this is one of the major advantages of considering organizations as social fields concerning the problem of self-organization. With respect to Bourdieu agents cannot maximize an abstract utility function regardless of a) the objective structure of positions which they occupy in the field, b) the logic, and c) the regularities of the field. The basis of this argument is the term habitus. The habitus of an agent is defined as a set of dispositions to specific ways of perception, thinking, and to perform actions. An agent is only capable to take a position because these dispositions acquired in a specific field enables it to perceive its specific chances and to act according to the objective possibilities available in the social field (for more details see [19]). The concept of habitus illustrates that human action is not an instantaneous reaction to immediate stimuli. Note that social reality does not only exist in social fields. It also exists in the habitus of agents. Bourdieu’s theory exhibits the role of the habitus as a necessary intermediate between the social structure of forces and the social action in social fields [18].

Regarding these basic assumptions of Bourdieu’s theory, organizations cannot be reduced to their formal structures. Their social structures have to be considered as cultural and political construction of dominant and dominated agents. Some agents are dominating according to their property and practical use of powerful resources like economic, cultural, social, and symbolic capital. Therefore, social structures are formed by relations of power whereby dominant agents like incumbents aim to reproduce their preeminent position over challengers and dominated agents which themselves try to conquer higher positions in the organizational distribution of power and authority [17].

To improve their position in the field, the agents need to play the game according to its own logic. Within the logic of organizations, members need to conceive means to carry out decisions and actions in the name of the whole to perform joint actions or tasks [20]. Formal structures can be viewed as such means. In this context, formal structures might be i) an object some agents want to change in favor of their interests, ii) a kind of capital or resources some agents use in favor of their interests, or iii) constraints according to which agents may act in a conform way because conformity is beneficial to them. The way how agents carry out joint actions can be considered as belonging to the logic of an organization as well. Organizations may differ as agents interact either in a cooperative, in a competitive or in a authoritarian way (c.f. for example [21]). For this reason, agents of different organizational forms use in our model different mechanisms to interact. For example, agents use the mechanism of gift exchange rather than economic exchange or authority, if they are members of an organizational form with a more cooperative character (see Section 3.2). Within the theory of Bourdieu the mechanism of gift exchange plays an important role for the creation of trust (cf. [17]: 191-202). Trust is needed to prevent opportunism especially in organizations in which self-interested agents are meant to work cooperatively together and in which neither authority constrains self-interested agents nor economic exchange guarantees efficient coordination.

Moreover, organizations are corporate agents which are embedded into macro-fields (e.g. the economic field) of the society. This means that these organizational agents are competing with other corporate agents in these meta-fields, trying to improve their objective position. As macro-social fields are sources of practice they constrain agents: Organizations need to cope with the regularities of the field and to act according to the logic of a social field as they would not be able to act without the structures of the macro-field (e.g. economic organizations need to make profits, accept the institution of market, or cope with legal regulations). Organizations do not have a habitus like individuals. Thus, an organization needs individuals with practical sense for representing the organization as a whole by means of social delegation (we will later need to this notion in Section 3.1). An organization is not only a corporate agent but a social field in itself in which agents are competing and trying to improve their positions. The individuals might accumulate capital (to achieve a better position) for themselves by improving the position of their organization in the macro-social field. Due to their socially structured interests based on their habitus, the individual agents will conceive "strategies" [18] how to reach a better position for themselves within the organization and probably for the organization within the macro-field. In summary, as long as the agents are interested in participating in the game, they build, reproduce, and change the organization in a self-interested way. Thus they try to improve their positions in the field using bounded social rationality [22]. The resulting structures and regularities may not be an optimal allocation of the interests or "utility" of every agent, but they enable joint action where decentralized mechanisms fail.

Organizations are not only social fields that appear, are reproduced and changed by the actions of a quantity of self-interested agents. To found an organization, member agents need to empower at least one individual agent to act for the whole [20]. Bourdieu developed a concept of social delegation, which is beneficial to the analysis of organizations. Bourdieu points out that it is necessary for the formation of a group or an organization to delegate a representative, which is empowered to speak for the organization to make the organization visible to the social environment [23]. Even though the delegate may abuse its position and its power in an opportunistic manner to improve or hold its own position within the organization, social delegation is an advantage. If all member agents of an organization would simply act according to their individual preferences, the organization as a whole might be unable to act and carry out joint action. It would be nothing more than a crowd of individuals. In this sense, social delegation may constitute a hierarchy between a quantity of agents and is a mechanism that enables coordination by authority. Therefore, this mechanism is necessary to provide the means that agents could found organizations and to enable self-organization in MAS. In our model we introduce social delegation in addition to the mechanism task delegation for this reason (see Section 3.1).

Last but not least, the determination of the boundaries between an organization and the rest of the social world (e.g. market interactions) is a necessary condition to transform a crowd of agents into an organization [20]. Within earlier organization theory there has been an analytical separation of markets and hierarchies (i.e. organizations). Since the seventies of the past century organizational networks became an important organizational form. Therefore, the determination of limits between formal organizations and market relations between organization became difficult [21]. These networks may not be completely economically and legally integrated, often they are only partially integrated by contract in order to persue a specific shared interest (e.g. a jointly fabricated product). We suggest to define these mixed forms as well as organizations, if they are not only bound by informal interactions, but have a formal structure as well (due to contracts, partially legally, economically integration, social delegation). With reference to Bourdieu, who remarks that the "limits of a field are situated at the point where the effects of the field cease to exist" (cf. [18]), the boundaries of these mixed forms are at that point, where e.g. a member of a network produces other products on his own independently from the network.

After we marked the basic insights of habitus-field-theory for self-organization of organizations as autonomous fields and corporate agents we have to point out the usability of this argumentation for the development of self-organizing MAS. In our view, we have to take into account the following aspects:

- From a sociological point of view self-organization of sociality (practice) is not a mechanistic but dynamic process. It emerges from the correspondence between the dispositions (habitus) of an agent to perceive, think and act and a social structure (field).
- Fields are self-organizing, emergent social entities. They show an objective structure of relations between social positions, a game-like character, and regularities which persist independently from single agents intentions and goals. Without these structures of the field agents are unable to act. On the other hand, only if agents are willing and able to act on the positions they have occupied, social practice is possible (for more details see [22, 17,24]). Fields are autonomous as far as the structures and regularities of a field are getting changed by agents attempting to improve their position within the logic of the field. Fields are self-organizing, not least, because the boundaries of a field are getting dynamically determined within the field itself by social struggles [18].

- Thus, if we consider the agent as the force behind the generation, change and reproduction of any social field we have to take into account that agents act in a self-interested way within a field as they try to improve their social positions.
- The self-interest of an agent cannot be represented by an utility function which remains identic in any situation. According to Bourdieu the rationality of an agent is socially bounded, i.e. it depends on its social position within a field in relation to the positions of other agents, the regularities and logic of a field, and on the situation.
- We consider the creation of organizations by the interactions of single agents as a very important example for self-organizing processes. Therefore, our concept of organization as an autonomous field and corporate agent is an example for self-organizing MAS. Any organization can be regarded as a social field, but not any field is as formally structured as an organization.
- Even though the basic characteristic of organizations is that they are formally structured social entities to carry out joint action, they have to be considered as autonomous fields and corporate agents in the sense of Bourdieu.

In the following section we will introduce FORM as a framework to transform these aspects into a model and design pattern of MAS.

3 FORM – A Framework for Self-Organization and Robustness in Multiagent Systems

In this section we present the *Framework for self-Organization and Robustness in Multiagent systems (FORM),* which is based on central issues of Bourdieu's theory (i.e. social field, capital, gift exchange, and social delegation) as described above. *FORM* is also founded on empirical sociological research on the genesis of social forms of organization (network building) and social structure in the field of transportation and logistics [24]. *FORM* is motivated by the close connection between robustness and self-organization in certain scenarios (for details on our view on the term "robustness" see [25]).

3.1 The Matrix of Delegation – A Grammar for MAS Organization

Recent work on delegation (see e.g. [26] for an extensive treatment), has shown that delegation is a complex concept highly relevant in multiagent systems, especially in semi-open systems. The mechanism of delegation makes it possible to pass on tasks (e.g. creating a plan for a certain goal, extracting information) to other individuals and furthermore, allows specialization of these individuals for certain tasks (functional differentiation and role performance). As we pointed out in Section 2.2, representing groups or teams is also an essential mechanism in situations which are dealing with social processes of organization, coordination and structuring. Following the concept of social delegation of Bourdieu, we distinguish two types of delegation: task delegation and social delegation. We call the procedure of appointing an agent as representative for a group of agents *social delegation.*

The activity of social delegation (representation) is in many respects different from performing tasks as described previously. For example it involves a possibly long-termed dependency between delegate and represented agent, and the fact that another agent speaks for the represented agent may incur commitments in the future, that are not under control of the represented agent. Social delegation is more concerned with the delegate performing a certain role, than with producing a specified product. In holonic terms, representation is the job of the head, which can also be distributed according to a set of tasks to different agents. Just like fat trees (multiple bypasses to critical communication channels) in massive parallel computing, distributing the task of communicating to the outside is able to resolve bottlenecks. This makes social delegation a principle action in the context of flexible holons and provides the basic functionality for self-organization and decentralized control.

Thus, we believe it is justified to differentiate two types of delegation: task delegation, which is the delegation of (autistic, non-social) goals to be achieved and social delegation, which does not consist of creating a solution or a product but in representing a set of agents. Both types of delegation are essential for organizations, as they rely on becoming independent from particular individuals through task and social delegation.

Given the two types of delegation, it remains to explain how the action of delegation is performed. We observe four distinct mechanisms for delegation:
(i) Economic exchange is a standard mode in markets: the delegate is being paid for doing the delegated task or representation. In economic exchange, a good or task is exchanged for money, while the involved parties assume that the value of both is of appropriate similarity.
(ii) Gift exchange, as an important mechanism in the sociology of Bourdieu ([17]: 191-202), denotes the mutually deliberate deviation from the economic exchange in a market situation. The motivation for the gift exchange is the expectation of either reciprocation or the refusal of reciprocation. Both are indications to the involved parties about the state of their relationship. This kind of exchange entails risk, trust, and the possibility of conflicts (continually no reciprocation) and the need for an explicit management of relationships in the agent. The aim of this mechanism is to accumulate strength in a relationship that may pay off in the future.
(iii) Authority is a well known mechanism, it represents the method of organization used in distributed problem solving. It implies a non-cyclic set of power relationships between agents, along which delegation is performed. However, in our framework authority relationships are not determined during design time, but the result of an agent deciding during runtime to give up autonomy and allow another agent to exert power. This corresponds to the notion of Scott who defines authority as *legitimate* power [27].
(iv) Another well-known mechanism is voting, whereby a number of equals determines one of them to be the delegate by some voting mechanism (majority, two thirds, etc.). Description of the mandate (permissions and obligations) and the particular circumstances of the voting mechanism (registering of candidates, quorum) are integral parts of the operational description of this mechanism and must be accessible to all participants.

In summary, we presented two modes of delegation and four mechanisms for performing each mode. Interestingly, all four mechanisms work for both modes of delegation and the combination of mode and mechanism spans a two-dimensional matrix. Theoretically, every combination of mode and mechanism is possible in multiagent organization: for example, economic exchange can be used for social delegation as well as for task delegation. Possibly this set of mechanisms is not complete, however, many mechanisms occurring in human organizations that seem not be covered here, are combinations of the described mechanisms.

3.2 The Spectrum of Organization

We will now describe seven different types of holonic organization for MAS in the order of increasing coupling between agents as shown in Figure 1. The types of organization are based on the framework of holonic MAS introduced in Section 2.1.

Modeling organizations of agents requires identifying schemes of joining first. In the following we describe five forms of organizations as ideal types we derived from empirical case studies as well as from organization literature about the initiation and emergence of organizations composed of formerly autonomous companies in the field of transportation and logistics [24].[1] The case studies were based on our sociological concept and the matrix of delegation. The forms we are presenting here are typical forms of organization occurring in human society. They differ significantly in the mechanisms generally used for task delegation and social delegation as well as membership limitations, profit distribution and the number of possible representatives. The difference between the forms is due to the positions of the formerly autonomous companies in the field, their strategies, the services they are supplying, etc. Therefore, we believe that these organizational forms are adequate for modeling (self-) organization in MAS using Bourdieu's theory.

In our presentation we will proceed from the most autonomous form of coordination along organizational types where agents partially give up autonomy up to a stage where they even surrender identity and merge into a single new agent[2] and focus on the differences in order to avoid redundancy where forms exhibit similarities. The agents may shift from market interaction to an organizational type where they give up only little autonomy to a form where they give up more autonomy and vice versa during runtime. The criterion for giving up autonomy is the volume of tasks they delegate among themselves in order to execute a complex task. The concluding summary of this section will then give a synopsis of all organizational types.

For modeling each of the types of organization we specify regularities, formal structures and the logic of these fields, stating what the organization's member agents are allowed to do, what they are obliged to do, and what they are forbidden to do. The delegation matrix provides the concepts for describing the interaction between agents.

Single, Autonomous Agents. This form of coordination is not of practical relevance but rather the theoretical starting point, with fully uncoupled agents. All agents that provide services do not interact with each other to accomplish their tasks, the only interaction taking place is between providers and customers.

[1] Note that in our current model single agents merging into an organization represent companies and not individuals. These companies are corporate agents (organizations) consisting of individuals. For future work the holonic approach allows to model these companies also as holons consisting of a number of single agents representing individuals.

[2] A discussion of the relation of this work to the concept of adjustable autonomy can be found in [28].

Example (see Figure 5(a)): A set of provider agents is not performing delegation of any type.

Market. At the very beginning, before agents choose an organizational type in a sociological sense for the first time, they are initiating relationships within the market. According to our empirical case studies, companies that jointly set up new organizations (e.g. virtual enterprises or alliances) or integrate other companies by acquisitions (e.g. group) usually already maintained trade relations with the companies joining the new organization. Sometimes, the enterprises already assigned orders to each other before (e.g. sub-contracting). However, these trade relations do not explicitly incur commitments in the future, as the transactions are based on one-time agreements of trade. Nevertheless, companies often assign orders to the same enterprise repeatedly, if they were satisfied with the quality and the price of the service. Still, these relationships are not formally structured. Thus, maintaining trade relations with one company does not imply that orders could not be delegated to any other enterprise. The companies do not appear jointly in the market, i.e. they do not necessarily need a joint representative.

Therefore, agents engage in our model in *task delegation* based on *economic exchange*. This means they exchange tasks and some kind of utility. This implies that agents build up relationships but not organizations in a sociological sense. Interaction is short-termed, based solely on the economic reasoning of the current interaction and aimed at increasing profit or keeping costs low, respectively. Coupling between agents is defined solely by economic exchange and agents can be members of many holons at the same time. The provider agent that redelegates parts of a task acts as the holon head for this specific task.

Example (see Figure 5(b)): provider agents are redelegating tasks in a market by economic exchange.

For determining the best bidder for task delegation online we use the *holonic contract-net with confirmation protocol (HCNCP,* see Figure 2*)*, which extends the *contract-net with confirmation protocol (CNCP)* (both: cf. [29]) to the holonic case and caters for efficient recursive or cascading application of the protocol. This is the protocol used by agents in the market, the virtual enterprise, and the alliance. The improvement for

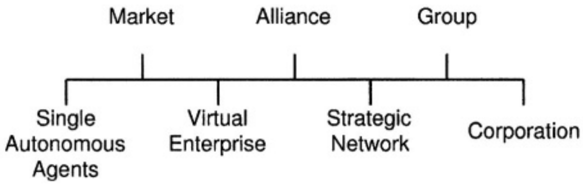


Fig. 1. Seven types of holon organization that are defined in analogy to the sociology of economic organizations. The types are arranged on a spectrum according to the intensity of coupling between participating agents. Note that the terms of this figure are used in the technical sense of describing types of organization and are not to be confused with the terms used in other contexts.

task assignment lies in the fact that agents only need to make a commitment about their resources when actually getting a task assigned (for a more detailed discussion see [30]).

Virtual Enterprise. The virtual enterprise is a loosely coupled set of participants organizing (possibly short-termed) to merge their core competences in order to produce a specific product not in the portfolio of any single agent. The strategies pursued with choosing this organizational type by self-interested agents (e.g. companies as corporate agents) is to reduce their investment costs by using the specific resources (e.g. social, economic, and cultural capital) of other companies in order to create one joint product. They choose this organizational type only for this specific product. Therefore, they can be members in other organizational types which produce other products and make use of their resources not allocated by the product of the virtual enterprise to make profit. It is possibly the stage of initiating tighter coupling between the participants [31,32].

The relationships between the companies are only to a small degree formally structured, not necessarily by contract. Therefore, the creation of trust among the members by gift exchange is very important to prevent opportunism. Although the participants remain economically and legally autonomous, they present themselves as a single company in the market throughout a company name, joint logo, etc., but only with regard to the specific product. Any company joining the virtual enterprise can act as representative of the whole. The model of this organizational type introduces longer termed *social delegation* that is specific to a single type of composed task. However, agents are still loosely coupled, every agent in the virtual enterprise holon can accept tasks from outside the holon and act for this task as the head agent. If it cannot solve the task by itself, it will then query other agents of the holon first for assistance. The mechanisms used here are *economic exchange,* and *gift exchange.* The role of gift exchange here is to be able to strengthen relationships to pave the way for tighter organizational types.

Example (see Figure 5(c)): Agents C, D, and E form the holonic structure of a virtual enterprise where C and D both act as heads, as they are both accepting tasks from provider agents outside the holon. C is also redelegating parts of its task to agents D and E inside the holon.

Alliance. An alliance as an organizational type is different to the virtual enterprise in that it is manifested by a long term contract among the participants and involves closer cooperation [33]. The relationships between the companies are formalized at least by contract, which is the result of negotiation between the different involved companies. Alliances are only to a degree economically and legally integrated. Therefore, the profit distribution is regulated for all internal transactions ex ante. Alliances are founded to create at least one new product. As the companies are only partially integrated they are usually supplying other products apart from the alliance. Thus, they are generally allowed to join other organizational types apart from those which produce the same product as the alliance. As most of the alliances are in some way legally integrated they need to appoint at least one CEO (representative). According to legal requirements, this is usually done by voting. Typically all member companies are equal in vote. Regarding strategic decision making processes the member organizations have equal rights as well. The interorganizational relationships within the network are rather cooperative than competitive or authoritarian. The representation of the alliance incurs valuable reputation and contact to customer agents implies (economic) power. Quitting of one of the agents with many customer contacts may cause loss to the organization, as customers may prefer to interact with the provider agent they already are acquainted with, no matter in which organization it is in. To decrease the incentive to join the alliance solely for this purpose and for the stability of the organization, a focal participant, who is, due to his already powerful position, not reliant on this increase in reputation, is appointed by *social delegation* through *voting* to represent the alliance. The profit is distributed among the head (representative) and all body agents necessary for performing the task by using *economic exchange* and *gift exchange.* However on creation of the alliance agents agree on a ratio (which is in our case fixed by the designer) that describes how profit is split between the head agent and the body agents that are involved in performing the task.

Example (see Figure 5(d)): Agents C, D, and E form an alliance. D is no longer allowed to act as a head. D has to forward any incoming tasks to its head C, essentially task delegation will be established only between the single head and the customer. In this case C even keeps the task, but it might also have redelegated it, depending on its own available resources.

Strategic Network. The strategic network differs from the alliance in that the relationships between the member organizations are rather authoritarian than cooperative or egalitarian. Typically, one company, which takes a high position within the field (domain) due to its economic strength, acts as focal or "hub" organization. This means that this enterprise coordinates the transactions between the members and takes the position of the representative. Nevertheless, a strategic network is not completely economically and legally integrated. The member organizations still remain autonomous, but they depend economically on the network to a great extent as most of their transactions take place within this organizational type. Therefore, memberships in multiple organizations are possible but unlikely. Strategic networks are founded to gain or sustain a competitive advantage towards their competitors by concentrating resources and exploiting synergies. According to our case studies the member organizations typically provide similar but not identical services which are combined or advanced to a high-quality service by the network, (i.e. in our model into a specific new product). Thus, a strategic network allows for reliably providing an enlarged portfolio. It is more reliable than the previous types of organizations, as by contract the focal participant has to a limited extent power over the actions of other participants [34]. *Authority* is introduced as mechanism for task delegation and agreed upon by contract. Also by contract the agents agree that profit is split by regulation as before in the alliance. In this organizational type anticipated coordination is demonstrated by the body agent's obligation to announce its cost function when joining the organization, as opposed to the previous types of organization where the head agent needed to request this information for every task. As a consequence, it is possible to use a shorter protocol than the HCNCP: the *direction with confirmation protocol (DCP),* see Figure 3.

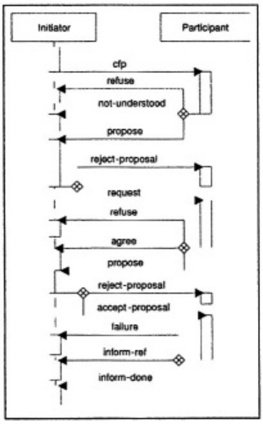


Fig. 2. The Holonic Contract-Net with Confirmation Protocol (HCNCP) which is used as the default protocol for efficient and cascading task assignment.

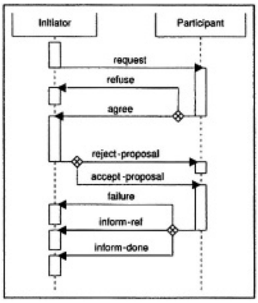


Fig.3. The Direction with Confirmation Protocol (DCP) which is used in the strategic network, making use of the knowledge of the initiator to reduce the number of messages required for task assignment.

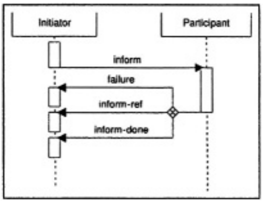


Fig.4. The Direction Protocol (DP) which is used in the group, making use of the knowledge of the initiator to further reduce the number of messages required for task assignment.

Example (see Figure 5(e)): Agents can participate in several strategic networks, each depicted by a circle around agents. In this case agent D is involved in two networks and receives payment (and tasks) from two heads. In contrast to the previous stages, it does not have the choice to negotiate about tasks as they are delegated by authority.

Group. A group (of companies) is different from a strategic network in that it is economically and legally integrated either by integration into a holding, acquisitions by a parent company or by a control agreement. Typically, the social structure of a group represents a bureaucracy or hierarchy with the most powerful company (parent company) at the top. All forms of practice in a group of companies are strictly regulated by a contract which is not the result of peer negotiation between the involved companies. Therefore, every part is only allowed to be member of this organization and not to be involved with any other [35]. Simple members of a group are typical restricted in many ways: All products created by a group company are handled as products of the whole group. No group company (apart from the top company) is allowed to build relationships to other external companies or customers on their own. Only the whole group is able to transact with customers. The profit distribution is centralized. No single company (again: apart from the top company) is allowed and able to regulate their internal and external economic transactions autonomously.

Concerning the modeling of a group, the relationship between the different parts of a group enacted by task delegation through *authority* is similar to that of the strategic network, but the consequence of the single membership restriction is that the head is informed about all tasks of each body agent. This means that the body agents (subordinate parts of the group) give up most of their autonomy to the head of the group. Messages to confirm that the agent can do a task are not necessary and we can use the shorter direction protocol (DP, see Figure 4). The downside for the head agent is that it is required to guarantee financial support, no matter how many orders can be acquired. Economic exchange is regulated by the constituting contract, gift exchange is not required as the relationship is also defined in the contract.

Example (see Figure 5(f)): Body agents are assigned tasks by authority as in the strategic network but must decide for one group membership. Agent D, which was part of two strategic networks now can only be part of one group.

Corporation. Merging of the agents with the loss of separation between the agents finally is the end of the spectrum: all agents provide their knowledge and resources for the creation of a single new agent. The merging of agents has been treated in technical terms for production systems for example by Ishida et al. [10].

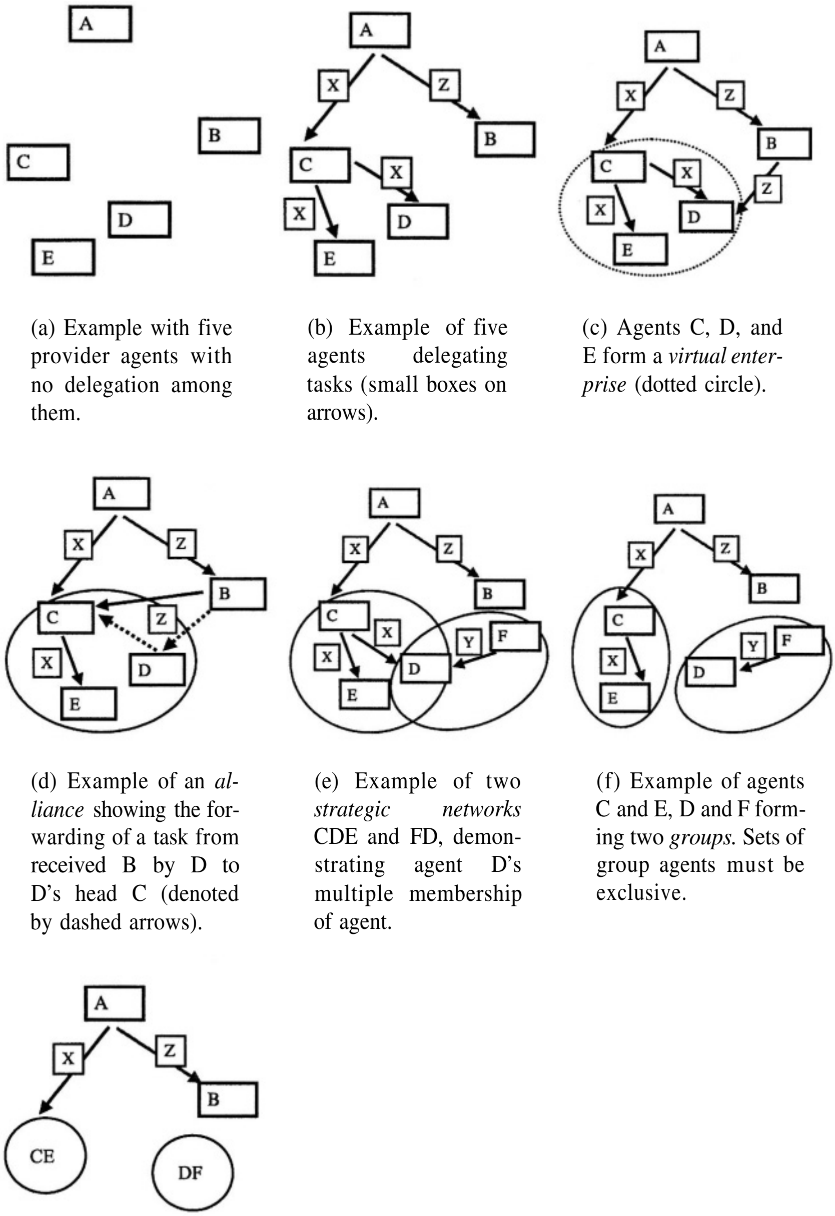
Example (see Figure 5(g)): Body agents C and D, E and F have merged into two new agents.

Summary. Although this discussion gives the impression that the spectrum is the process of several agents merging through different intermediate stages into one agent, it is a process that depends on the current situation of all participating agents. Each individual agent will choose, depending on the situation in the MAS, whether it is in its interest to proceed with the process. As each organizational type has advantages and disadvantages, it may well be, that a transition is not beneficial in the light of the current market situation. It is also worth noting that each stage of the organization here builds on earlier stages, and introduces new restrictions. Therefore, we can speak of a total ordering of the organizational type and hence, a *spectrum* of organizational types.

While the previous sections described the addition of new institutions or the removal of an institution when proceeding to a new organizational type, Table 1 gives a synopsis on the organizational types. In short, we will address the properties mentioned and relate them to the institutions defined previously. The table concentrates on the five organizational types that are practically relevant (omitting the non-interacting agents and the single agent) and lists six different properties. The properties TD and SD stand for the mechanisms that are used for task delegation and social delegation as described by the matrix of delegation (see Section 3.1).

Property M (membership limitations) can have the value “limitation on product”, which means that the agent is free to chose other organizations to join, as long as they were not formed to perform tasks with the same set of resource types. M can also denote that there is no limitation (as in the market) or mean that an agent is only allowed to be member of one single organization (as with the group).

Profit distribution (PD) can be performed on a per task basis using economic exchange or economic exchange and gift exchange. Other possibilities imply that during the formation phase of the organization agents agree on how profit is split between head



(a) Example with five provider agents with no delegation among them.

(b) Example of five agents delegating tasks (small boxes on arrows).

(c) Agents C, D, and E form a *virtual enterprise* (dotted circle).

(d) Example of an *alliance* showing the forwarding of a task from received B by D to D's head C (denoted by dashed arrows).

(e) Example of two *strategic networks* CDE and FD, demonstrating agent D's multiple membership of agent.

(f) Example of agents C and E, D and F forming two *groups*. Sets of group agents must be exclusive.

(g) Example for *corporation:* body agents C and D, E and F have merged into two new agents.

Fig. 5. Overview on the seven types of holonic organization.

Table 1. Overview of the seven types of holonic organization. Rows specify for each type the mechanism for task delegation (TD), social delegation (SD), the membership limitations (M), the mode of profit distribution (PD), the role of the holon head (HH) and the protocol used for task assignment (P).

Property	Market	Virtual Enterprise	Alliance	Strategic Network	Group
TD	Ec Ex	Ec/Gift Ex	Ec/Gift Ex	Authority	Authority
SD	Ec Ex	Ec/Gift Ex	Voting	Authority	Authority
M	No Limitation	Limitation on Product	Limitation on Product	Limitation on Product	Exclusive Membership
PD	Ec Ex	Ec/Gift Ex	Regulation	Regulation	Fixed Income
HH	One/All	All	One	One	One
P	HCNCP	HCNCP	HCNCP	DCP	DP

and body agents ("regulation", e.g. 10:90, 20:80 etc) or that a "fixed income" is being paid from the head to the body agents regardless of the number of tasks performed (in this case, variable costs are paid by the head plus a fixed income chosen by the designer).

The role or number of the holon heads is described by property HH. In the virtual enterprise all agents can receive incoming tasks and redistribute them. For the market the situation is to some extent up to interpretation: Although there is only one task per holon and only one agent communicating and coordinating for this holon, all agents in the system are allowed to accept tasks and then engage in coordination and communication. With the other organizational types the property is very crisp again, as all other forms allow only a single point of access to the outside. Depending on the organizational type we chose to model the use of three different protocols (property P) as depicted in Figures 2, 3, and 4.

4 Conclusions

We presented our *Framework for Self-Organization and Robustness in Multiagent Systems (FORM)* to describe organization and self-organization for task-assignment MAS. This framework is inspired by theoretical considerations using the habitus-field-theory of Pierre Bourdieu and it was supported by empirical sociological research on organizational forms within the domain of transportation and logistics. Our contribution to the concept of organization applies central terms of Bourdieu's theory (e.g., social field, capital, gift exchange, social delegation) and focuses on both organizations as social fields as well as corporate agents. Our framework is based on a matrix of delegation that serves as a grammar for MAS organization using gift exchange and social delegation as basic mechanisms. *FORM* is an extensive description for the creation of MAS that makes use of *holons,* a well established concept for designing organizational structures. In contrast to previous concepts of social organization, *FORM* offers a new approach to the modeling of autonomy and self-organization in task-assignment. The advantages of a sociological concept of organization on a higher level of social aggregation were shown by using a new kind of social mechanism (gift exchange) and a new model of delegation (splitting it into task delegation and social delegation). As a result, *FORM* overcomes the dilemma of either modeling self-organization in the sense of a simple coordination of the interactions of single agents, or to model MAS-organizations statically. In contrast to the increased complexity of the organizational model, our model is not static and allows membership in multiple organizations (unless explicitly forbidden) to build different organizational structures concurrently and dynamically adapt them to the environment. Precisely in this sense, *FORM* allows to model self-organization in MAS. *FORM* is currently being implemented as a testbed for MAS and we investigate a decision model that drives the agents' choice of an organizational form from the framework. In future work we will investigate the power of *FORM* by conducting experiments to evaluate the effect of self-organization on performance and robustness of task-assignment in MAS.

References

1. Bond, A.H., Gasser, L.: Readings in Distributed Artificial Intelligence. Morgan Kaufmann (1988)
2. Mason, C., Johnson, R., Searfus, R., Lager, D., Canales, T.: A seismic event analyzer for nuclear test ban treaty verification. In: Proceedings of the Third International Conference on Applications of Artificial Intelligence in Engineering. (1988)
3. O'Hare, G.M.P., Jennings, N.R.: Foundations of Distributed Artificial Intelligence. John Wiley and Sons (1996)
4. Weiß, G.: Adaptation and Learning in Multi-agent systems: Some remarks and a bibliography. In: Adaptation and Learning in Multi-agent systems. Springer (1996)
5. Jennings, N.: Agent-based computing: Promise and perils. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99). (1999) 1429–1436
6. Fischer, K., Schillo, M., Siekmann, J.: Holonic multiagent systems: A foundation for the organization of multiagent systems. In: Proceedings of the First International Conference on Applications of Holonic and Multiagent Systems (HoloMAS'03). (In print)
7. Turner, P.J., Jennings, N.R.: Improving the scalability of multi-agent systems. In: Proceedings of the First International Workshop on Infrastructure for Scalable Multiagent Systems, Barcelona, Spain, June 2000. Lecture Notes in Artificial Intelligence, vol. 1887. Springer-Verlag (2001)246–262
8. So, Y., Durfee, E.H.: Designing tree-structured organizations for computational agents. Computational and Mathematical Organization Theory **2** (1996) 219–246
9. Axtell, R.: Effects of interaction topology and activation regime in several multi-agent systems. In: Multiagent-based simulation: Second International Workshop (MABS 2000). (2000) 33–48
10. Ishida, T., Gasser, L., Yokoo, M.: Organization self design of production systems. IEEE Transactions on Knowledge and Data Engineering **4(2)** (1992) 123–134
11. Carley, K.M.: On the evolution of social and organizational networks. Research in the Sociology of Organizations **16** (1999) 3–30
12. Jennings, N.R., Faratin, P., Norman, T.J., O'Brien, P., Odgers, B.: Autonomous agents for business process management. Int. Journal of Applied Artificial Intelligence **14** (2000) 145–189
13. Rosenschein, J., Zlotkin, G.: Rules of Encounter. The MIT Press, Cambridge, Mass. (1994)

14. Roth, G., Schwegler, H., eds.: Self-organizing Systems. An Interdisciplinary Approach. Campus, Frankfurt/New York (1981)
15. Fischer, K.: Holonic multiagent systems — theory and applications. In: Proceedings of the 9th Portuguese Conference on Progress in Artificial Intelligence (EPIA-99), LNAI Volume 1695. Springer-Verlag (1999) 34–48
16. Koestler, A.: The Ghost in the Machine. Hutchinson & Co, London (1967)
17. Bourdieu, P.: Pascalian Meditations. Polity Press, Cambridge, Eng., Oxford, Eng. (2000)
18. Bourdieu, P., Wacquant, L.: An Invitation to Reflexive Sociology. Polity Press, Chicago (1992)
19. Schillo, M., Zinnikus, I., Fischer, K.: Towards a theory of flexible holons: Modelling institutions for making multi-agent systems robust. In: 2nd Workshop on Norms and Institutions in MAS, Montreal, Canada (2001)
20. Argyris, C., Schön, D.A.: Die lernende Organisation. Klett-Cotta, Stuttgart (1999)
21. Powell, W.W.: Neither market nor hierarchy. network forms of organization. Research in Organizational Behavior **12** (1990) 295–336
22. Schillo, M., Fischer, K., Hillebrandt, F., Florian, M., Dederichs, A.: Bounded social rationality: Modelling self-organization and adaptation using habitus-field theory. In: Proceedings of the Workshop on Modelling Artificial Societies and Hybrid Organizations (MASHO) at ECAI 2000. (2000)
23. Bourdieu, P.: Delegation and political fetishism. Thesis Eleven **10/11** (1984/1985) 56–70
24. Dederichs, A.M., Florian, M.: Felder, Organisationen und Akteure-eine organisations soziologische Skizze. In: Bourdieus Theorie der Praxis. Erklärungskraft - Anwendung - Perspektiven. Opladen, Wiesbaden (2002) 69–96
25. Schillo, M., Bürckert, H., Fischer, K., Klusch, M.: Towards a definition of robustness for market-style open multi-agent systems. In: Proceedings of the Fifth International Conference on Autonomous Agents (AA'01), Montreal, Canada, ACM, New York (2001) 75–76
26. Castelfranchi, C., Falcone, R.: Towards a theory of delegation for agent-based systems. Robotics and Autonomous Systems **24** (1998) 141–157
27. Scott, W.R.: Organizations: Rational, Natural and Open Systems. Prentice Hall Inc., Englewood Cliffs, N.J. (1992)
28. Schillo, M.: Self-organization and adjustable autonomy: Two sides of the same medal? In Hexmoor, H., Falcone, R., eds.: Proceedings of the AAAI2002 Workshop on Autonomy, Delegation, and Control: From Inter-agent to Groups. (2002)
29. Schillo, M., Fischer, K., Knabe, T: The contract-net with confirmation protocol: An improved mechanism for task assignment. Technical Report TM-01-01, DFKI (2001)
30. Schillo, M., Kray, C., Fischer, K.: The Eager Bidder Problem:A Fundamental Problem of DAI and Selected Solutions. In: Proceedings of the First International Conference on Autonomous Agents and Multiagent Systems (AAMAS '02). (2002) 599–607
31. Kemmner, G., Gillessen, A.: Virtuelle Unternehmen. Physica-Verlag (2000)
32. Fischer, K., Müller, J.P., Heimig, I., Scheer, A.W.: Intelligent agents in virtual enterprises. In: Proc. of the First International Conference on Practical Applications of Intelligent Agents and Multi-Agent Technology (PAAM'96), London. (1996)
33. Gulati, R., Garguilo, M.: Where do interorganizational networks come from? American Journal of Sociology **104** (1999) 1439–1493
34. Jarillo, J.C.: On strategic networks. Strategic Management Journal **9** (1988) 31–41
35. Freichel, S.L.K.: Organisation von Logistikservice-Netzwerken. Erich Schmidt Verlag, Berlin (1992)

Social Organization in a Software Agent Community with a Non-zero-Sum Game Interaction Model

Matti A. Vanninen and John R. Rose

Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29208 USA
matti@sc.edu,rose@cse.sc.edu

Abstract. An artificial society of learning software agents employing preferential partner selection and the Iterated Prisoner's Dilemma as a means of social interaction is described. The IPD has been used extensively in modeling interaction and the development of cooperation in communities, but existing research has focused predomi0nantly on identifying successful strategics in tournaments and ecological simulation. This simulation models an alternative scenario, one in which the members of the community are fixed, but are endowed with a personality and the ability to develop preferences among the other members of the community. It was predicted that such agents would learn to associate with agents which they judge to be favorable, and that the community would achieve a higher net efficiency than it would if interactions occurred randomly. Communities of highly selective agents were found to perform consistently better than non-selective ones.

1 Introduction

1.1 A Social Dilemma

Individuals in a community are motivated by conflicting goals of maximizing personal benefit and maximizing the communal good. Classical economic theory, exemplified by Adam Smith's "invisible hand", suggests a society benefits from egoism among its members, but this naïve assumption often does not hold in real life. On the contrary, frequently actions which are clearly advantageous for an individual are equally disadvantageous for the community, and thus paradoxically for the individual himself insofar as he is a member of the community. The Tragedy of the Commons [1] and Voter's Paradox [2] are typical examples. It is such social dilemmas that form the bulk of any public policy debate, and the need for government in the first place.

The Prisoner's Dilemma is a well known instance of this type of problem. There are two players, each of whom must choose to cooperate or defect without knowing what the other will choose. Payoffs are assigned to each player for each of the four possible outcomes, and each player seeks to maximize his own payoff. The particular values for the temptation to unilaterally defect (T), reward for mutual cooperation (R), punishment of mutual defection (P), and sucker's payoff (S) may vary, so long as T>R>P>S, which ensures that the dominant choice and optimal choice are different.

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 176–188, 2004.

The dominant choice is to defect, since defecting will result in a better outcome regardless of the adversary's choice. However, mutual cooperation is better than mutual defection for both participants, and is the globally optimal outcome.

The Iterated Prisoner's Dilemma is an extension of the Prisoner's Dilemma in which the participants play the game repeatedly, and the final payoff is the cumulative sum of individual payoffs. Each iteration is resolved like the PD, and each player knows the results of all previous iterations. In the IPD, a further condition is that `2R>T+S` and `2P<T+S`, since otherwise the dilemma can be escaped by alternating cooperation and defection. While participating in an IPD, a person employs a strategy which determines the choice made for a given iteration. The strategy is a mapping from the game history to an action. Strategies can be categorized as pure and mixed, where the mapping in pure strategies is deterministic, that is, the same decision will always be made given the same history, whereas the mapping in a mixed strategy is randomly selected from a distribution determined by the history.

1.2 Cooperation and the IPD

The Iterated Prisoner's Dilemma game was chosen for this purpose because it is a ubiquitous tool in the study of social processes and the emergence of cooperation in societies. Despite its superficial simplicity, the IPD elegantly captures the essential problem of decision making with incomplete information. Like the tragedy of the commons and many other social dilemmas, the IPD is a non-zero sum game. Victory for one party does not imply an equal and opposite defeat of the other. Instead it is possible for all participants to achieve varying degrees of success or failure, and therefore a strategy which is most disadvantageous for other participants is not necessarily optimal. Maximizing the misery of the opponent through constant defection is not nearly the best possible outcome in the IPD; mutual cooperation is much better. However, the game is paradoxically designed to reward defection over cooperation for individuals. How does cooperation exist, when defection is the rationally superior choice?

The development of cooperation in a system which rewards exploitation was initially studied by Robert Axelrod. Axelrod's method, which has since been frequently duplicated, was to host a competition in which IPD strategies are paired in a round robin tournament and ranked by overall number of victories, identifying strategies which perform well in a diverse environment. The experiment was extended through ecological simulation, in which underperforming strategies are removed in subsequent rounds of the tournament, identifying robust strategies which perform well against other strong strategies. From his observations, Axelrod extrapolated the following qualities of a successful IPD strategy [3]:

1. Don't be envious.
2. Don't be the first to defect.
3. Reciprocate both cooperation and defection.
4. Don't be too clever.

Axelrod's rules were based predominantly on the success of Tit-for-Tat, a strategy which mimics the opponent's previous move after initially cooperating, and which has proven to be a highly robust strategy. The third point is particularly counterintuitive, since logically defection is superior to cooperation in any circumstance. Heuristics such as these, which contradict logic but lead to a collectively and individually superior outcome, are thought to be the basis for the emergence of cooperation in systems which do not seem to support it. Only the last rule has been seriously challenged. Axelrod noted that in his tournaments strategies which attempted to be overly clever tended to fail since the other participant is unable to predict their decisions. Gradual, a complex strategy, has since then been conjectured to be superior to Tit-for-Tat [4].

The success of any strategy in a tournament is always dependent on its environment [5]. For example, among Defectors, Tit-for-Tat will always lose, and among hard mistrustful strategies, Tit-for-Two-Tats fares considerably better than TFT. Ecological simulations reveal the robustness of a strategy, which is generally a better indicator of its fitness than its success in any particular setting, and are hence favored over tournaments

2 Project Description

2.1 A New Model

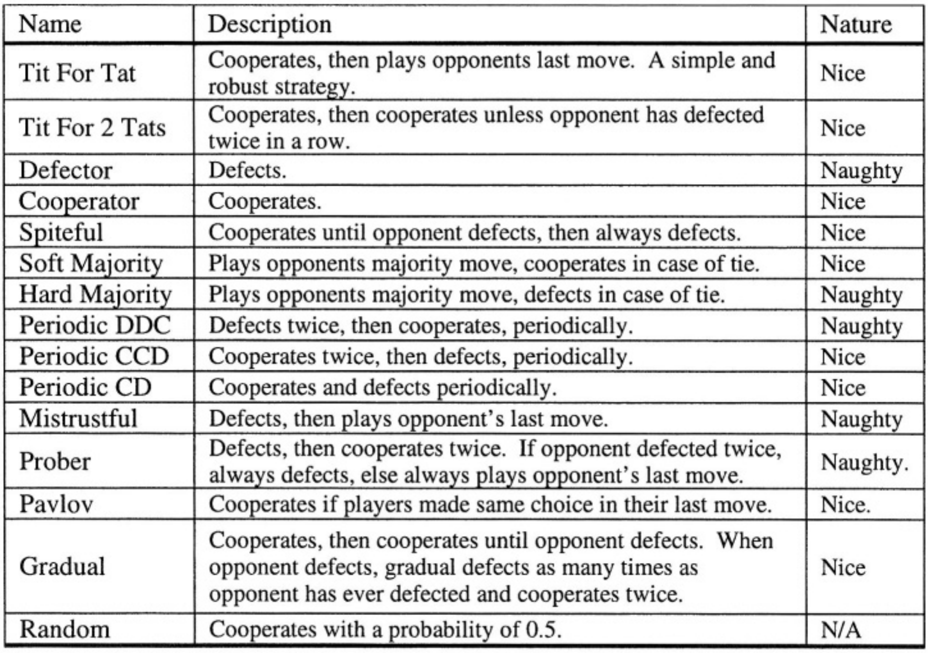
Existing research on the IPD has typically focused on identifying optimal strategies and variations of the game, such as stochastic or non-fixed payoff games (Ashlock [6], Axelrod [7], Eriksson [8], etc.) The robustness of a strategy is usually determined by evaluating performance in a round robin tournament or an evolving simulation. However, this is unsuitable for modeling the short behavior of human populations since such populations do not evolve ecologically.

As an example, consider the people in a neighborhood, each of whom has the personal goal of improving his or her happiness. An obvious means to this end is to develop friendships and social networks. The process interaction among two (or more, but for the purposes of this study we limit ourselves to two) persons can be considered a game in which an increase in well being is victory and a decrease a defeat. It is intuitively clear that this is a non-zero sum game since social interaction can be a mutually satisfying experience. Another possibility is that an incompatibility between persons causes each to be dissatisfied, or that one takes pleasure at the expense of the other. Of course if members of the community inherently value bilateral cooperation above all, determining the outcome of a given interaction is not particularly challenging. However, if exploitation is assumed to be more profitable than cooperation (i.e., people are greedy) and being exploited is assumed to be the worst possible outcome (i.e., people are jealous), the game is an instance of the Prisoner's Dilemma. An ecological model does not capture the fact that although the relationships in the community are dynamic, membership is largely static. The rate of change of people in the population is negligible compared to the rate at which friendships are made and broken, and people are unlikely to alter their personalities to adapt to their social climate. The ecological model also does not capture the values of the community. For example, if a base level of satisfaction for each member of the community is more important than extreme success for some at the cost of others, the optimal behavior of individual agents will not necessarily be to maximize personal happiness. These complexities require an alternate model, one where poorly performing strategies are not eliminated, but are able to redirect their energy toward identifying and interacting with more compatible partners.

2.2 Description and Implementation

The artificial society for this model was constructed using software agents. The tasks of the autonomous units that make up "Persons" in the society are fairly straightforward to encapsulate within the software agent paradigm, and since a large collection of tools is available for working with software agents, this was the approach used in implementation. The agents in the model are essentially collaborative learning agents, in that they act autonomously and react to their environment by segregating acquaintances according to favorability as they are encountered [9]. The simulation is performed as a number of rounds, in which each Person in the community has the opportunity to interact with one other Person. An interaction between a pair of Persons is called a conversation, and is modeled as 100 iterations of the IPD.

Table 1. IPD strategies available in the model



Name	Description	Nature
Tit For Tat	Cooperates, then plays opponents last move. A simple and robust strategy.	Nice
Tit For 2 Tats	Cooperates, then cooperates unless opponent has defected twice in a row.	Nice
Defector	Defects.	Naughty
Cooperator	Cooperates.	Nice
Spiteful	Cooperates until opponent defects, then always defects.	Nice
Soft Majority	Plays opponents majority move, cooperates in case of tie.	Nice
Hard Majority	Plays opponents majority move, defects in case of tie.	Naughty
Periodic DDC	Defects twice, then cooperates, periodically.	Naughty
Periodic CCD	Cooperates twice, then defects, periodically.	Nice
Periodic CD	Cooperates and defects periodically.	Nice
Mistrustful	Defects, then plays opponent's last move.	Naughty
Prober	Defects, then cooperates twice. If opponent defected twice, always defects, else always plays opponent's last move.	Naughty.
Pavlov	Cooperates if players made same choice in their last move.	Nice.
Gradual	Cooperates, then cooperates until opponent defects. When opponent defects, gradual defects as many times as opponent has ever defected and cooperates twice.	Nice
Random	Cooperates with a probability of 0.5.	N/A

Strategies. Fifteen strategies were chosen for this model as they represent a wide cross section of characteristics, and are mostly well known in the literature. Strategies which cooperate initially are described as "nice". Strategies which open with a defection are "naughty". As was empirically demonstrated by Axelrod and others, this distinction often has a dramatic impact on the overall performance of a strategy.

Personality Parameters. An essentially novel aspect of this model is the augmentation of each member of the community with a personality. The distinct patterns of behavior of individuals in a superficially homogenous society can be attributed to personal factors and values. The behavior of the agents in this model is controlled by four parameters in addition to the strategy. These were identified as shyness, selectivity, optimism, and envy, labeled in rough correspondence to actual personality characteristics (see Table 3 for definitions). Each Person exhibits each trait to some degree, ranging from 0 to 1. The parameter values are used as weights in decisions made during the conversation partner selection and evaluation phases. The objective of this model is not to provide an accurate simulation of human psychology, but to investigate the behavior of diverse agents in a society, and the personalities are intended only as analogues. Nonetheless, they are representative of some of the parameters which are involved in decision making among members of societies.

Development of the Simulation

Outline:

1. Interface announces start of simulation to Bulletin Board
2. Bulletin Board broadcasts start of round report to Persons
3. Persons use **shyness** attribute as a weight in a random decision to initiate or listen to a conversation
4. Initiating Persons use **optimism** attribute as a weight in deciding whether to converse with a known acquaintance or a random, possibly unknown Person. In the former case, **selectivity** attribute is used in deciding which acquaintance is chosen.
5. Persons announce their decision to the Bulletin Board.
6. Bulletin Board matches initiators to listeners, satisfying partner requests as well as possible, and announces pairings to Persons.
7. Persons hold conversations though 100 rounds of IPD, communicating via the Arbiter.
8. Persons adjust opinion of conversation partner using conversation score disparity and **envy** parameter.
9. Persons report conversation outcome to Bulletin Board and Interface.
10. When all conversations are complete, Bulletin Board initiates a new round.

Prior to the start of the simulation, the world will be populated by agents known as the Bulletin Board, Arbiter, Interface, and one or more Persons. The Interface does not participate in the proceedings of the simulation, and its functions will not be further discussed. The Bulletin Board and Arbiter are aware of each other, and the Persons are aware of the Bulletin Board and the Arbiter. As Persons enter the world, their first task is to report themselves to the Bulletin Board, and wait for acknowledgement. The Bulletin Board maintains a database of the names of all Persons it is aware of, along with the world state.

The simulation is initiated when the Bulletin Board is externally prompted (generally by the Interface). The Bulletin Board begins a round of conversations by sending a report to each Person, and waits for each Person to acknowledge the message. When a Person is informed that a new round of conversations is beginning, it first decides whether to take the role of a conversation Initiator or Listener. An Initiator may further select a preferred conversation partner from its database of acquaintances. The Person sends an acknowledgement to the Bulletin Board, and waits to be assigned a partner.

When all Persons have acknowledged and reported their role, the Bulletin Board creates pairings of Initiators and Listeners. First, the Bulletin Board attempts to satisfy the requests of those Initiators who requested a preferred partner. If the partner is unavailable, either because that Person did not elect to be a Listener or has already been assigned to another conversation, no partner is assigned to the Initiator. Once all possible preferences have been satisfied, the Bulletin Board assigns random partners to the remaining Initiators. Persons are informed of their partner as the pairings are made, and once the supply of Initiators or Listeners is exhausted, any remaining Persons are informed they will not be participating in this round.

Once an Initiator is informed of its partner, it informs the Arbiter of the new conversation. The Arbiter creates a table to record information about this conversation, and requests a decision from each participant, beginning the conversation. The arbiter may, and usually will, control multiple independent conversations at once.

Table 2. Summary of Agent Roles

Agent Type	Agent Role
Person	Represents an actor in the community. Attempts to maximize personal score through interaction with other Persons
Bulletin Board	Acts as a directory through which Persons discover one another. Satisfies requests by Persons to listen to or initiate conversations. Maintains global statistics
Arbiter	Acts as an impartial intermediary in a conversation among two Persons. Relays decisions between the conversing agents.
Interface	Reports global an individual statistics, and allows user to set parameter values.

The conversation is modeled as 100 iterations of the Iterated Prisoner's Dilemma game. The game proceeds with the Arbiter sending an information request to each Person for each iteration. The request contains the current status of the game (score of each side, and the previous move of each participant). The Persons use this information and their personality attributes to make a decision, and respond with their intent to cooperate or defect. The Arbiter notes the decision, which may be misinterpreted if the noise of the world is greater than zero, and sends a new request. After 100 iterations, the Arbiter informs the Persons that the game is over, and reports the final result. The Persons use the results and their personality attributes to evaluate their opponent, and add their partner to their database of acquaintances, or adjust their opinion of the partner if they are previously acquainted. The Initiator reports the end of the conversation to the Bulletin Board. Once all conversations have ended, the Bulletin Board initiates a new round.

Partner Selection. A Person's first task in a round of conversations is to decide whether it will be an initiator or listener in this round. The shyness attribute is employed in making this decision. If a Person chooses to be an Initiator, it must also decide whether to request a preferred partner. If the Initiator has no acquaintances, no preference can be made. Otherwise, the Initiator uses its optimism attribute. Even if no preference is made, the Bulletin Board may assign a partner with whom the Initiator is already acquainted.

If the Initiator has decided to prefer an acquaintance, it uses its selectivity attribute to determine which acquaintance is selected. The rule is:

```
If (selectivity == 1)
       Select favorite acquaintance
For each person
       If favorability <= 0
           prob = 0
       Else
           prob = favorability^(selectivity/(1-selectivity)
For each person
       Prob = prob/max_prob
```

This algorithm assigns a probability of selection for each person. If selectivity is 0, all persons have the same probability. If selectivity is 0.5, the probability of selecting a person is equal to the selector's opinion of that person, if selectivity is 1, the best friend will always be selected, etc. In other words, selectivity determines how important a person's opinion of another is in deciding whether to pursue a conversation.

Conversation Progression. When it is informed of a new conversation, the Arbiter requests each Person to make a decision in the Prisoner's Dilemma game. In the request, the Persons are informed of the current score of each participant, and the previous decision of each participant. The Arbiter does not maintain any history of the conversation, except for the current round number and score. The Persons make a decision based on the IPD strategy they are employing. Once the Arbiter has received both replies, it requests another round, etc., until 100 rounds have occurred. The Arbiter then informs each Person that the conversation has ended, and reports the final score.

Partner Evaluation. After a conversation has finished, each Person adjusts its opinion of its partner using the outcome of the game and the envy parameter. The rule is:

```
scoreWeight=myScore/optimalScore
advantageWeight=(myScore-hisScore)/optimalAdvantage
rating=advantageWeight*envy+scoreWeight*(1-envy)
favorability =(favorability*(numencounters+rating)/numencounters
```

The opinion is the average of all the ratings the person has made of his partner. The rating is based on the absolute score of the Person (the scoreWeight), and the margin of victory over its partner (advantageWeight). An envious person rates a partner based solely on its margin of victory without regard for the numerical result, whereas a tolerant (non-envious) Person will rate a partner based solely on his score independent of the partner. In this game, optimalScore and optimalAdvantage are both 500.

Noise. In addition to the personality traits and partner selection, this model attempts to accurately represent an actual community with the inclusion of a noisy environment. Error and miscommunication are represented with a noise parameter, which is the probability that a decision will be misinterpreted. This makes the progress of the simulation nondeterministic, and helps overcome a very important criticism of the Prisoner's Dilemma as a sociological tool: humans in reality do not actually act rationally. An alternative approach would have been to incorporate more mixed (stochastic) strategies. The noise parameter effectively transforms all strategies into mixed strategies, and hence the effect of each approach would be similar. Mixed strategies tend to be ineffective among deterministic ones in a deterministic environment, while a robust pure strategy should be able to compensate for miscommunication.

The noise parameter is handled internally by the Arbiter, which coordinates and judges conversations among the member agents in the community. All decisions relayed by the arbiter have a probability equal to the value of the noise parameter of being inverted. For example, if noise is 0.01, one percent of decisions to cooperate are interpreted as defections, and vice versa. A conversation is modeled as 100 rounds of the IPD, and one decision is made by each participant for each round, there will be on average two errors in a given conversation.

3 Experiments

Sandholm & Crites [10] have explored the success of learning strategies in a heterogeneous environment, but unlike their model, the agents in this society do not adapt their strategies but rather alter their partner selection preferences. Ashlock, Smucker, Stanley, and Tesfatsion explored a similar scenario, in which strategies express partner preference in an evolving society. They use an algorithm similar to the one in this model to select an interesting conversational partner, with the addition of a possibility to refuse a conversation request. However, they studied an evolving model, with the goal of discovering individual robust strategies, whereas the purpose of our model is to discover whether partner preference in a static society will lead to a better outcome for the entire community. Skyrms and Pemantle [11] also analyzed the behavior of a community with a non-zero sum game interaction model and preferential partner selection, but their method of partner evaluation is substantially different from ours. Macy & Flache [12] studied the evolution strategies of psychologically endowed agents participating in social dilemmas, but their research focused on the impact of aspirations and habituation on the search for a good equilibrium between two agents, whereas we investigate the impact of an altogether different set of traits on the development of relationships in a community.

The impact of the selectivity and envy attributes were explored experimentally. The parameter settings are summarized in Table 3. The value of selectivity was varied from 0.0 to 1.0 among a community of ten agents while maintaining all parameters except noise at 0.5. A random strategy was assigned to each strategy, and the mean score per round per agent after 250 rounds was recorded. The experiment was repeated forty times for each value of selectivity. A similar set of experiments was performed with varying values for the envy parameter. Examining a non-deterministic system was expected to give a better insight into real world behavior, as it allows for the possibility of miscommunication and other variables which the model does not directly account for. The experiments were thus repeated within a noisy environment.

Table 3. Summary of Personality Traits and Experimental Parameter Settings

Attribute	Function	Experimental values
Selectivity	Willingness to converse only with highly esteemed acquaintances	0 to 1, increments of 0.1 Envy parameter set to 0.5 when testing impact of selectivity.
Envy	Intolerance of a relatively poor outcome in a conversation.	0 to 1, increments of 0.1. Selectivity parameter set to 0.5
Optimism	Expectation that the outcome of a conversation with an unknown person will be positive.	Not tested. Value of 0.5 used in all tests.
Shyness	Inhibition against initiating conversation	Not tested. Value of 0.5 used in all tests.

The success of the trials was evaluated by comparing the average score per conversation in a community to the average score if the strategies were paired arbitrarily. The predicted mean score depends on the set of strategies used, and for the strategies in this study, it was found to be approximately 220 for a pure environment, and 215 with 5% noise. The raw score is not a useful indicator of performance, since the actual number of conversations an agent participated in is unknown. For example, a naïve strategy might obtain a high score because it is frequently selected as a partner by exploiters, and so it participates in a large number of conversations, even though its outcome in any one conversation is weak. Adding a cost to participate in a conversation would have been an alternate way to distinguish between strategies which perform well and ones which are active often.

4 Observations

A breakdown of the results of the experiments follows. The charts show the average score per conversation in the community for various parameter values. Each point is the mean score of 40 communities employing randomly selected strategies but endowed with the same set of personality parameters. Unless specifically noted, the parameter values were 0.5. The expected mean score for arbitrary pairings is denoted by a the dashed line. A score above the expected mean suggests the agents were able to adjust to their environment and form beneficial relationships. The payoffs for the game were T=5, R=3, P=1, S=0.

The correlation between selectivity and score is clear. Highly selective communities perform substantially better than those which are not selective. The correlation between envy and score is less obvious, but there is a downward trend as the value of envy increases. In both cases, the value of the parameter has little impact on the relative performance of the community if it is less than 0.5, implying it is effectively masked by the other parameters. Higher values dominate the other parameters, leading to a notable upturn for selectivity and a slight downturn for envy. The mean score per conversation per agent in the most highly selective community is approximately fifteen points higher than the mean score for arbitrarily paired strategies. The highly envious community scores approximately equal to the predicted mean. Thus selectivity is a beneficial parameter, whereas envy is not.

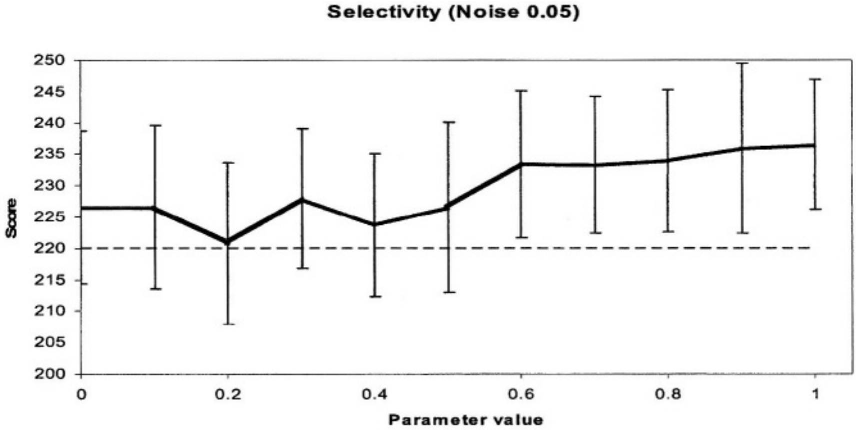


Fig. 1. Average score per conversation increases as value of selectivity increases, and is consistently higher than expected mean for arbitrary pairings (dashed line). Error is one standard deviation.

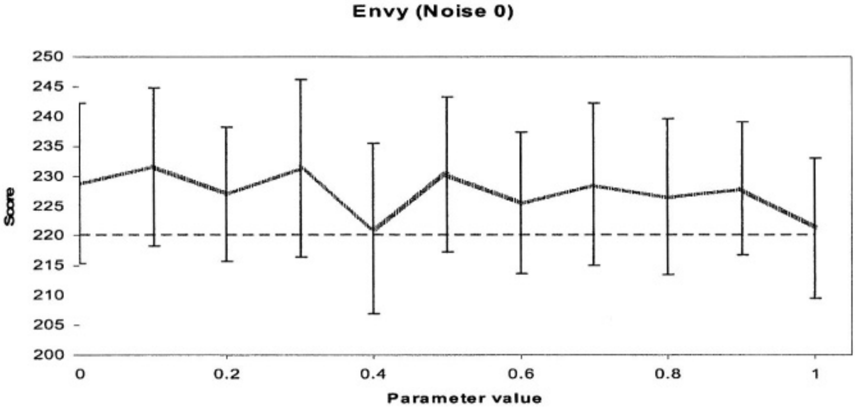


Fig. 2. Average score per conversation decreases as value of envy increases. Error is one standard deviation.

The correlation between selectivity and envy and score is less evident in Figures 3 and 4. Both plots follow a similar path, peaking at about 0.7. Note that the expected mean score in a noisy environment is lower than in the pure game since most of the participating strategies do not forgive unexpected defections. Altering the values of the personality parameters makes little difference in the relative performance of the communities. The clear upward trend for selective communities is not evident, nor is the downward trend for envious communities. The noise in the environment appears to undermine the ability of the agents to evaluate and select their partners.

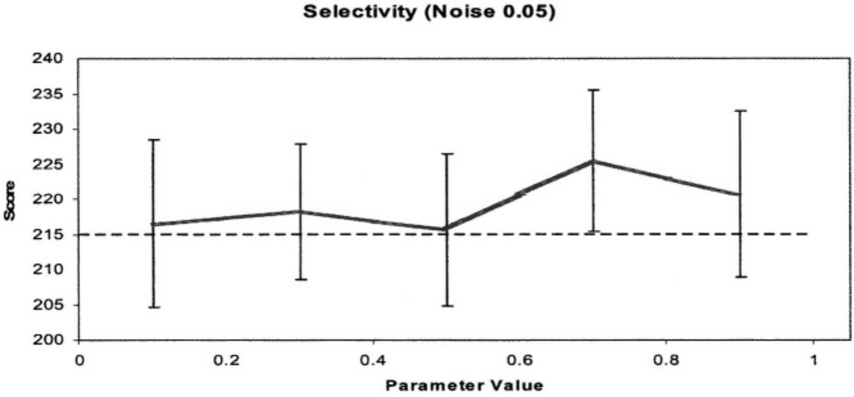


Fig. 3. Correlation between average score per conversation and selectivity is difficult to establish in noisy environment. Error is one standard deviation.

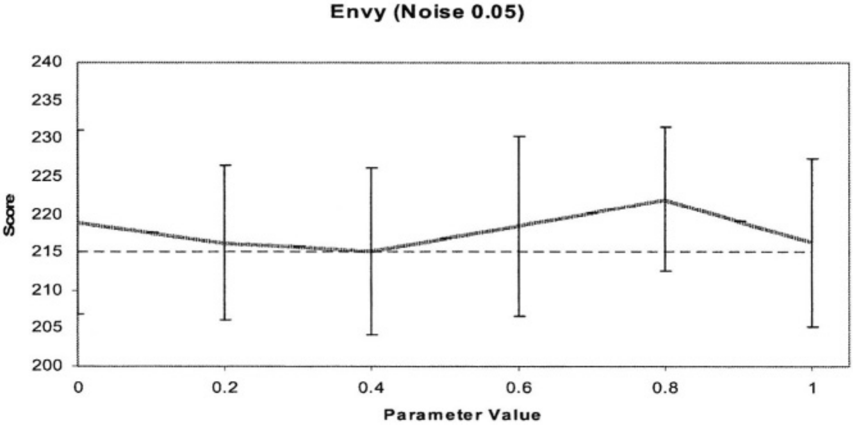


Fig. 4. Correlation between average score per conversation and envy is difficult to establish in noisy environment. Error is one standard deviation.

5 Conclusions

The strongest result in these experiments was the outstanding performance of selective communities. When given the opportunity to evaluate and select their partners, the agents developed beneficial (mutually or otherwise) relationships with other members of the community and performed better than predicted for random selection. The success of the agent society model suggests that including partner preference and personality variation in IPD-based communities is a good way to enhance the value and realism of the simulation.

The impact of the envy parameter is less clear. There does appear to be a downward trend in mean score per conversation as the value of the envy parameter increases. This was the initial prediction, and suggests that highly competitive individuals do not necessarily improve the net well being of a community. Initial data had implied an alternative view, in which the exploiters in an envious community are rewarded by the naïve strategies, and mutual defection, or laziness, which is the worst possible outcome, is discouraged more effectively than it is in tolerant communities. However, this broader dataset mirrors the more conventional attitude that competition is a behavioral opposite of cooperation.

That the ability of the agents to adapt to their environment is severely undermined by even a slight element of chance is also evident. In nearly all cases, agents which perform consistently better than average in a pure game perform unpredictably in a mixed environment. This suggests the patterns of interaction which develop to promote cooperation are fragile, and easily covered by noise. In all cases, the agents in a noisy environment substantially under perform those in a noise free environment, further supporting this theory. We believe the issue is with the particular strategies which were selected for the model. Many are pure and unforgiving. For example, the majorities, gradual, spiteful, and tit-for-tat inherently perform poorly in an even slightly noisy environment. Selecting some more forgiving strategies is recommended for any future studies. Increasing the size of the communities would also help alleviate this problem. With only ten agents, individual rogue strategies may destabilize the entire group. Spiteful, for example, performs atrociously with even a slight amount of noise, and the majorities, tit-for-tat, and gradual are not very tolerant of error. Their weaknesses as individuals may have partially caused the poor performance of the non-deterministic communities in general.

References

1. Hardin, G.: The Tragedy of the Commons. Science. 162 (1968) 1243-1248
2. Felkins, L.: The Voter's Paradox. http://www.magnolia.net/~leonf/sd/vp-brf.html (1995)
3. Axelrod, R.: The Evolution of Cooperation. BasicBooks New York (1984)
4. Beaufils, B., Delahaye, J., Mathieu, P.: Complete Classes of Strategies for the Classical Iterated Prisoner's Dilemma. http://www.lifl.fr/IPD/references/ from_lifl/ep98/html (1998)
5. Campbell, R.: Background for the Uninitiated. In: Campbell, R. and Sowden, L. (eds.): Paradoxes of Rationality and Cooperation. U. of British Columbia Press Vancouver (1985)
6. Ashlock, D. et al.: Preferential Partner Selection in an Evolutionary Study of Prisoner's Dilemma. BioSystems, 37(1,2) (1996) 99–125
7. Axelrod, R.: The Complexity of Cooperation. Princeton Univ. Press (1997)
8. Eriksson, A., Lindgren, K.: Evolution of strategies in repeated stochastic games with full information of the payoff matrix. Spector, Lee et al (eds.)Proceedings of the Genetic and Evolutionary Computation Conference - GECCO-2001, Morgan Kaufmann Publishers (2001)
9. Software Agents: An Overview. BT ISR Agent Research http://193.113.209.147/projects/agents/publish/papers/review2.htm
10. Sandholm, T.W and Crites, R. H.: Multiagent reinforcement learning in the iterated prisoner's dilemma. BioSystems, 37(1,2) (1996) 147–166

11. Skyrms, B., Pemantle, R.: A Dynamic Model of Social Network Formation. Proceedings of the National Academy of Sciences of the USA 97 (2000) 9340–934
12. Macy, M.W., A. Flache.: Learning Dynamics in Social Dilemmas. Proceedings of the National Academy of Sciences USA. 99 (2002) 7229–36.

Emotion: Theoretical Investigations and Implications for Artificial Social Aggregates

Christian von Scheve[1] and Daniel Moldt[2]

[1] University of Hamburg, Institute of Sociology,
Allende-Platz 1, D-20146 Hamburg, Germany
xscheve@informatik.uni-hamburg.de
[2] University of Hamburg, Computer Science Department,
Vogt-Koelln-Str. 30, D-22527 Hamburg, Germany
moldt@informatik.uni-hamburg.de

Abstract. One of the most pressing issues in the social sciences and in distributed artificial intelligence research is the micro-macro link that is the question of how individual action and social structure are interrelated. Besides others disciplines, sociological research has identified emotion as being a possible key component in this link. Unfortunately, sociological theories in question remain relatively basic, and do not refer to emotion research from other disciplines. We show that emotion theories and models from cognitive science, psychology, neuroscience, and computer science constitute a valuable, if not mandatory foundation for sociological issues in emotion research. We therefore present an integrated view on emotion. The goal is to relate specific micro-macro aspects of emotion theory with general sociological theories of societal structuration. This issue is briefly discussed in the context of an exemplifying multi-agent architecture.

1 Introduction

This paper analyses the interrelation of emotion and social structures in natural and artificial social aggregates. One of the key problems, both in distributed artificial intelligence and in the social sciences is the micro-macro link, i.e. how individual action is related to social structures and vice versa [71]. In this article, we argue that emotion plays a major role in this linkage. It is hypothesized that emotion is capable of "absorbing" structured physical and mental environments and of "impinging" them on an individual's information processing architecture. Over and above that, subjects continually recreate these structures by means of emotionally biased behavior of diverse kinds.

We will briefly outline functional basics of emotion in individuals and also focus the *link between* two or more socially interacting subjects and how social order is supposed to emerge from these interactions. To do this, we draw upon a wide range of research results from various disciplines such as psychology, neuroscience, sociology, and computer science.

Computer science as the only mentioned discipline not directly concerned with research on natural emotions is considered both, an enabler and profiteer of our investigations. Computer science can enable research in this field by providing techni-

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 189–209, 2004.

ques to model, depict, and simulate complex systems, processes, and interdependencies while probably profiting in many ways from new insights into the social dimensions of emotion which are ideally presented in a formal, agent-based model [12, p. IX].

The article is structured as follows: In the second section we present notions and methodologies and specify the goals we pursue. Then we briefly summarize the latest developments in research on emotions regarding the disciplines in question. The fourth section illustrates our approach to integrate emotion theories from diverse disciplines, focusing on the social world as one cause and consequence of emotion. In the fifth section we make suggestions on how our theoretical findings could be combined with aspects of Pierre Bourdieu's and Norbert Elias's social theories which have already been modeled by means of agent-oriented Petri nets. Finally, we draw conclusions and give an outlook on future work.

2 Means and Methods

This section describes our research goals, the methodological approach we pursue, and defines important terms.

2.1 Goals and Methodology

Our research goal is threefold:

1. To gain new insights into the *social* causes and consequences of human emotion by combining research results from those disciplines concerned with analyses on the micro-level (e.g. neuroscience, psychology) with results and *open questions* found within the social sciences, traditionally concerned with questions of social aggregation (macro-level analysis) [60].
2. In computer science – especially in the fields of human-computer interaction (HCI) and DAI – there is an increasing need for theories of emotion that explicitly account for large scale social dimensions and that can easily be related to existing approaches. Therefore, we strive for a theory that (a) explains the social structural components of emotion as well as their dynamics, and (b) is formulated in a way that makes it useful for computational models.
3. It is not a new insight that computer science and the social sciences could mutually benefit from broader foundations for agent- and multi-agent system-concepts [26]. Malsch [54] has coined this endeavor "socionics". In this respect, cooperation of computer science and sociology could support the construction and analysis of large scale (social) systems – but emotion in a sociological interpretation has not yet been covered in an appropriate way. This is probably due Weber's [81] and Parsons' [64] conceptualizations of action: "Under the aegis of this conceptualization, emotion was regarded as not only irrational but pre-modern: such views became sociological conventions" [3, p. 16].

Future software systems will on the one hand involve many human participants and on the other hand they will (probably) be designed following the MAS-paradigm. Questions with respect to the interaction of these human and artificial actors are numerous. In this context, emotions being generated, shaped, and transformed in / by these systems and emotions generating, shaping, and transforming these systems need to be investigated. For these questions we provide a theoretical background as well as basic requirements for an emotion-based MAS-architecture.

Why, one may ask from a social scientist's point of view, co-operate with computer science? Where are the benefits for social theory? When conducting emotion research in an interdisciplinary way, combining micro- and macro-level analyses, we consider an actor-centered approach to be the most suitable. Benefits then result from three observations:

1. Because neuroscientific and psychological emotion research is strongly actor-centered, and
2. sociological emotion research dealing with macro causes and effects is also largely actor-centered [33,43,78]. Moreover, there is considerable consensus in sociology that macro-phenomena can in some cases be traced down to micro-acts and instances [15,45].
3. Because in computer science *agents* are an increasingly popular and promising concept. Conceptually they can be understood as a technological counterpart to human actors. To fulfill the pretensions of autonomy, intelligence, mobility, sociability or even emotionality, aspects of the human or animal cognitive system are interpreted as a model for agents' formal reasoning and behavior generation (decision-making, plan-generation, action-selection) [82].

Thus, agents and multi-agent systems are ideally suited to simulate and possibly validate theories that employ the actor as a central concept. Furthermore, due to methodological and theoretical heterogeneity in the distinct disciplines which conduct research on emotion, a conceptual framework is needed that is capable of incorporating and interfacing different theories and concepts. Computational, agent-based models and modeling languages are designed to describe and depict complex systems of various kinds in formal, operational semantics. In this respect, a plea for more profound and formal models in the social sciences, especially in sociology, has been made by [77]. In our opinion, the concept of an emotion system is of such complexity and analysis thereof can fundamentally profit from formal, computational models.

Considering emotion theory, our method is a qualitative-heuristic analysis according to [44]. It is not our intention to build a completely new theory of emotion. Instead, we present first steps towards an integrative view on the diverse and broad theoretical perspectives. Qualitative-heuristic analysis is a means to discover "blind spots" in a specific theory. Although much work is currently done [30,32,41], it is our conviction that most "blind spots" in emotion theory today still can be found at higher levels of social aggregation. Many questions concerning social aggregates could be answered by interfacing and extending existing theories.

2.2 Notions

Much work in computer and cognitive science has been done examining emotion in isolated entities and in (dyadic) social interactions (see the next section for an overview). However, little research has been carried out scrutinizing emotion in the context of larger social aggregates and comprising the role of social structural implications.

Social aggregates (or units, if one likes) like groups, teams, communities, societies or organizations, are considered to be forms of social interaction which are mutually, repeatedly, and orderly carried out by a specific, although possibly dynamic number of individuals. Social aggregates are not necessarily required to be coherent in time and space – they may exist independently of physical presence or time disparities. Natural social aggregates are made up by the interactions of (human) actors, whereas artificial social aggregates require artificial agents (e.g. BDI-agents [31]) to interact with each other (e.g. acting on behalf of one and the same user/client or sharing a common goal).

Social aggregates have specific qualities of diverse kinds such as norms, rules, laws, rites, institutions, etc. For any individual within a social aggregate it is important to have either implicit or explicit knowledge about these qualities to be able to act in relation to these qualities. They may constrain or enlarge actors' options for action and facilitate interactions among actors. These qualities are not objects of the physical world, rather they are "mental objects" of individuals within a social aggregate, and they are internalized by learning or socialization [17,45]. Depending on how actors act in relation to these qualities, the structure of a social aggregate remains more or less stable. Also, joint actions, coalitions, and co-operations require participating actors to act in congruence with norms or rules.

In view of emotion, we are foremost interested in their functional components and will neglect phenomenological, physiological, and related issues here. It is of primary interest, how emotion influences (neuro)cognition and vice versa, and – what is specific to our approach – how this relation affects and is affected by societal conditions. We perform a functional-conceptual analysis in order to resolve questions mentioned above and refer to [39, p. 203] for more detail in this respect. Therefore, specific emotions such as fear, anger, sadness, joy or the like are not accounted for, neither are "social" emotions distinguished from "non-social" emotions.

We define emotion as a state or process that mediates, influences, and is influenced by social, perceptual, physiological, and higher cognitive capabilities of an entity. They are "functional, organized responses to environmental demands that prepare and motivate the person to cope with the adaptational implications of those demands" [74, p. 36]. In human actors, emotion consciously or unconsciously facilitates information processing, verbal and non-verbal communication, social behavior, action selection, decision-making, etc. Emotions have phylogenetic and ontogenetic components, of which the latter are of primary interest here. That means we will analyze components which are alterable during an individual's lifetime (runtime), e.g. by socialization, adaptation and learning.

Choosing this definition and understanding of emotion should not be (in view of computer scientific models) considered superficial – instead it facilitates simulation by omitting possible questions of subjective experience or embodiment.

Although computers may probably never *subjectively experience* emotion (at least in the near future and in a phenomenological sense), it is no question that computers can have "special states that correspond functionally to emotions in organisms" [61].

Having thus clarified our goals we proceed illustrating theoretical and computer-scientific research on emotion to make our goals clear more precisely and to show the urgent need for such an approach.

3 Emotion in Human Actors and Artificial Agents

This section resumes trends and results in theoretic, empiric as well as in computer scientific emotion research. Although, clear-cut distinctions between the different disciplines cannot always be made, we subsume disciplines according to our research goals under "sociology" and "cognitive sciences". With this distinction the focus is either on macro- or micro-level analyses. "Cognitive science" encompasses disciplines such as cognitive and social neuroscience, and cognitive psychology, whereas "sociology" focuses sociological and social psychological research.

3.1 Sociology

The recently established field of sociological emotion research has – in our opinion – not yet fully realized the importance of emotion for social life and social phenomena. Although elaborated and original work exists (e.g. [42,78]), the majority of sociological research tends to neglect important findings from psychology and/or neuroscience despite the fact that there are many valuable connections made with social issues in these lines of research. The sociology of emotion has a long time struggled with intradisciplinary rows between so called positivist and social constructionist positions [43]. By now, it seems that radical constructionist positions [72] have been abandoned and the moderate positivist position is widely accepted.

Most important contributions from sociology (also regarding problems in psychology and artificial intelligence) emerge from areas dealing with *inter-* rather than intrapersonal aspects of emotion, an issue that has somewhat been neglected in psychology [13, p. 212]. Inseparable from this are aggregational (macro) causes and consequences of emotion, a topic sociologists have made valuable contributions to. Kemper for example argues that emotions result from social relationships which are in turn characterized by social status and power [43, p. 344]. Social structures, i.e. vertical stratifications on the macro-level, are made up by the distribution of the social resources status and power amongst individuals. Thus, social structure and emotion influence each other reciprocally.

Collins [16] on the other hand, illustrates that the exchange of "emotional energy" in social interactions facilitates societal structure generation. According to Collins, individuals have an inherent drive to keep up a certain level of "emotional energy" and therefore steadily seek interactions that provide a gain of emotional energy and avoid those that cause a loss.

Although these are valuable contributions toward understanding the relation between individual behavior and social structures, almost all approaches from sociology lack concrete evidence, specifications, and testable models. Collins's [16] concept of "emotional energy" for example is hardly defined at all and remains very vague throughout his explanations.

One step toward more precise and specific models is made by the newly emerging sub-discipline neurosociology that combines neurological evidence obtained by (functional) magnetic resonance imaging (MRI/fMRI) or other techniques with sociological theories of interaction and structuration [25]. Unfortunately, most works presented in that volume lack cognitive foundations, so that an important part of the emotion process is once again not accounted for.

Therefore, to thoroughly understand the social causes and consequences of emotion, an integrative perspective is needed that comprises and interrelates the social, cognitive, and neurological dimensions of emotion.

3.2 Cognitive Sciences

Without a doubt, the most comprehensive research on emotion has come from psychology, with a strong emphasis on cognitive and social psychological theories, whereas the predominant perspective has been intraindividual [55, p. 202]. The results of the diverse works are too extensive even to be summarized here, instead we will focus and very briefly introduce *central topics* and *conceptual models* on which considerable consensus has been achieved.

One of the most prominent and central issues in psychology is the cognition-emotion interrelation that is lead by questions on how emotions influence thoughts and how thoughts influence emotions. There is unchallenged evidence that emotional states decisively affect human cognitive performance, such as problem solving, learning, memory formation, attention, judgment, decision-making, etc. [6,14,24,38].

Despite these results derived from experimental psychology that mainly scrutinize the effects of feelings on cognition, there are elaborated theoretical approaches to emotion that shed light on the question how cognition generates and regulates emotion. Departing from the discussion between [50] and [83] on the question, if cognition is at all involved in emotion generation, there now seems to be wider agreement on *appraisal theory* as one conceptual approach [63]. Basically, appraisal theorists assume that cognitive evaluation of external stimuli generates a subjective meaning on which emotions are based.

The "primacy of affect" [83] within this concept can e.g. be explained by refined information processing theories, such as Leventhal and Scherer's [52] perceptual processing theory that divides perceptual processing into hard-wired sensory-motor, internalized schematic, and inferential conceptual processing as a basis of appraisal.

However, evidence from cognitive neuroscience suggests that emotions *can indeed* occur without any (higher) cognitive involvement [51]. Furthermore, as stressed by other researchers, emotional reactions and their consequences for cognition and overt behavior are often (socially) conditioned and unconscious [18]. Attention to the unconscious has been largely disregarded in sociological (emotion)theory since Max Weber's [81] definition of social action as *intentional* behavior. Although, many prominent works describe mechanisms of structuration whilst tacitly assuming the existence and effectiveness of unconscious determinants of social action (e.g. [28]). The significance of unconscious activity of the emotion system lies within those substantial influences of emotion on cognitive activity which are not accessible by conscious deliberation and do not enter awareness. These mechanisms give conditioning and socialization a whole new meaning because as long as actors are not aware of them, they can hardly be intentionally altered or regulated. Thus, they emphasize the significance of the functional components of emotion over those of subjective feelings.

3.3 Computational Models

Computational models of emotion seek to capture and synthesize functional and expressive components of emotion in the first place; subjective feelings are far beyond what is currently achievable, discussed or even desirable.

The emerging field of "affective computing" is defined by [66] as "computing that relates to, arises from or deliberately influences emotions". In broad terms, the field can be subdivided into efforts to capture and model emotional user states, to synthesize emotions in AI systems for optimized reasoning or decision-making capabilities or to build emotionally expressive systems for richer interactions. Many of the up to date approaches prefer agent oriented systems design, either to make use of methodological advantages or to realize better implementation of emotion theories.

Researchers in the area of affective computing consider emotions to be a crucial part of overall intelligent behavior or as [56, p. 163] stated: "The question is not whether intelligent machines can have emotions, but whether machines can be intelligent without any emotions". Therefore, in order to build systems that are capable of exhibiting intelligent behavior, computational models of emotion are needed which fit into currently used techniques from the field of artificial intelligence.

By now, research conducted on the various aspects of affective computing is focused on cognitive and recently also on social components of emotion, whereas the social dimension is analyzed mainly in view of dyadic agent-human or agent-agent interactions. There are efforts to increase performance and efficiency by means of emotional heuristics [70,80], to improve interactions [5,40,65] or to analyze the role of emotion in artificial minds [73].

Still largely detached from affective computing and related AI-techniques is a continued trend towards distributed AI (DAI) systems. DAI systems rely on the assumption that intelligence is not primarily a matter of isolated entities but rather a question of socially interacting entities [19,2]. Besides the intelligence debate, there are endeavors to transfer the qualities of primate (including human) or animal societies, i.e. natural social aggregates, to computational systems. These qualities are robustness, failure-tolerance, adaptivity and autopoiesis [12]. In this respect, multi-agent systems, i.e. artificial social aggregates, are currently the most promising methodology [22].

To achieve the above mentioned qualities of natural social aggregates, research is actually focused on social phenomena such as coordination, cohesion, cooperation, trust, commitment, and the like. Only recently and very partially, emotions have been considered to be an important part of these phenomena and of global system behavior [1,11,29,75]. However, important findings from sociology dealing specifically with social structural aspects – which are of great interest here – have been largely neglected so far.

We think that there is an enormous potential for computational, especially distributed (and possibly affective) systems, in marrying the neurological, cognitive, and social (sociological) components of emotion; a position that will be illustrated in more detail in the following section.

4 An Integrative Approach

This section describes our integrated approach to analyze the social components of emotion and how they can be found in each of the disciplines addressed above. We first illustrate the influential forces of sociality in different domains and then depict in detail how they are interrelated.

4.1 Sociality as a Common Issue

What has been disregarded in many theories of emotion so far is the fact, that the social worlds individuals are located in are more than a mere collection/aggregation of social agents inhabiting this world. Social systems have specific qualities that emerge from interactions taking place within this system – but these qualities and their causes often cannot be traced back to an individual agent. Nevertheless, these qualities are a major source of influence on any agent's biological, cognitive, and emotion system – in other words: on the determinants of an individual's overall behavior, be it overt (external) or covert (internal). These qualities do not only affect, as it has been assumed for a long time, an individual's "social conventional" actions acquired by learning and priming, but more profoundly also the very basis of an individual's information processing architecture.

We will show in which way we consider emotion to be one key component in the micro-macro link, that is how emotions are directly influenced by social phenomena and via intermediate neural and cognitive pathways, and how emotions and their neural and cognitive underpinnings work concerted to maintain or alter social structural qualities.

As we have briefly illustrated in the preceding section, the various approaches each shed light on specific components of human emotions, such as emotion and social structures, emotion and neural correlates, emotion-cognition relations, and the synthesizing of emotions. Since we aim at finding a stable link between macro-aggregates and micro-acts, we first have to examine if and in which way that, what is widely accepted to be "social", possibly affects the components of an individual's information processing architecture relevant for the emotions.

Second, we have to analyze to what extent social processes and structures influence that what operates on this information processing architecture, namely cognitive activity and mental representations.

Third, it is of interest how these internal mechanisms become involved in communication and social interaction, how they are expressed, interpreted, and judged, and how they become (through bodily or verbal manifestation) part of a social environment.

4.2 Social Neuroscience

As briefly illustrated before, findings from neuroscience suggest that "rational" decision-making based on "pure reason" or "formal logic" is – at least in cognitive tasks serving socially oriented purposes or personal future outcomes – hardly achievable [18, p. 170-3]. To explain the possibility of successful and quick decision-making in such tasks, Damasio introduces the somatic-marker hypothesis. Somatic-markers can be thought of as a biasing device that (unconsciously) assists human deliberation in reducing alternative options by emotionally marking appropriate (positive) and inappropriate (negative) options.

Damasio goes on to explain that somatic-markers are not predefined or hard-wired in the emotional system, rather they are acquired during (early) socialization and education by "connecting specific classes of stimuli with specific classes of somatic state" [18, p. 177]. Thus, defective or highly erroneous human decision-making in socially oriented tasks, such as in certain types of sociopathy, are at least partly traceable back to maladjusted social development, unless pathological conditions are indicated.

Thus, somatic-markers are a neural and therefore hardly correctable or avoidable means by which behavioral regularities in a social environment, particularly during primary socialization (parents, peers, and friends), can be impinged upon an individual's information processing system. By provoking specific emotional reactions to specific classes of stimuli (real or imagined), a certain form of behavior tendency, also of "non-emotional" character, is promoted. These behavior and decision-making tendencies, we presume, roughly resemble characteristics of prevailing socially shared cognitions and common emotional reactions in the social environment, i.e. the social aggregate, an individual is socialized in. We certainly do not deny individual differences in emotional reactions – emotion to a great extent is what makes us "individual". We also firmly acknowledge subjective interpretations of the social world, which precede any establishment of somatic-markers. But, as we will argue in the following section, initial and supposedly subjective interpretations are also biased by social forces.

Furthermore, there is evidence from the social neurosciences that the very basis of cognitive and emotional activity, the physiological structure and development of certain brain regions (individual's information processing architectures), is affected by social environmental conditions (see [10] for an overview). As [9,27] have argued, socio-cultural factors play an important role in how the brain organizes and selects incoming information, e.g. from the sensory cortices. That means, brains are transducers, they "[..] *change* environmental information (to which the organism could not otherwise respond) into physiological processes that can be received and processed into something humanly meaningful" [25, p. 159, italics original]. Tredway and associates state, that "critical to the formation of a well developed limbic system are *healthy affective* interactions, especially during infancy and early years" [76, p. 110, italics added].

Without further investigation of the latter issue here, we conclude that social aggregational qualities (the social environment) impinge specific modes of biological development, of information processing and (emotional) behavior upon individuals. As long as these forms of behavior are of no pathological kind, we assume that they serve to maintain the structures of social aggregates that originally built them.

The results from neuroscientific investigation set in relation to research efforts in the sociology of emotion (see above) suggest a picture of micro-macro linkage that is fundamentally based on the neural underpinnings of emotion. What has been examined and described by sociological emotion researchers such as [16,42] as well as by sociologists like Elias [21] or Bourdieu [7] could find its more "evidential" foundations in the affective and social neurosciences. We will refer to this possibility in more detail in the following sections.

4.3 Social Cognition

As we have sketched above, there is an unquestioned interrelation between emotion and cognition, and because of their tight connection, both seem only to be conceptually and possibly anatomically, but not functionally separable. In the process of behavior, there is no zero-level emotion or cognition state, unless in some pathological cases. Thus, behavior is neither solely cognition-driven nor solely emotion-driven.

The preceding section has shown in how far social environments may shape emotional responses regardless of *actual* higher cognitive activity operating on working memory. This section examines in which way cognitive activity that is relevant for emotion processes depends on social environmental conditions. In doing this, we refer to the models mentioned before, namely cognitive activity in the appraisal process and in different modes of information processing. Central to this endeavor are the concepts of social cognition and distributed cognition [23].

Besides the aspects of externalization and temporal distribution of cognition, the social distribution of cognition seems to be most relevant for the emotions [62, p. 82]. Socially distributed cognition describes the distribution of cognitive activity (on a specific task) among different individuals. This is either to achieve goals that could not otherwise be accomplished individually (complex or large task-domain, insufficient knowledge) and requires cooperation and coordination, or to overcome deficiencies of individual cognition, such as biases in social judgments [8].

In order to synchronize cognitive activity on a specific task, individuals probably have to adapt their cognitive style to the requirements of their peer-group. Especially from a developmental perspective and since "many, perhaps most, human activities involve socially distributed cognition" [62, p. 83], one can assume that the prevailing or most successful cognitive style within a specific social aggregate is presumably adopted by other individuals up to a certain degree.

Social cognition, on the other hand, is cognitive activity that selects, interprets, and uses social information to make judgments and decisions about the social world. Central concepts are *schemas* and *scripts.* Schemas are a collection of related beliefs individuals use to organize their knowledge about the social world. Upon perception of a certain class of stimuli, one categorizes other persons (stereotypes) or the roles they perform to fit a known schema. Actions and further inferences are often based on a schema rather than on what is actually perceived, i.e. on raw data [4]. Scripts are schemas about events and situations and involve action and behavior strategies.

Although scripts and schemas help to behave according to norms and rules or to act and decide quickly, they are a major source of erroneous behavior (in situations and encounters deviating from standard every-day situations), since possibly valuable information is filtered and not accounted for. Schemas and scripts are based upon past experiences; they are socially learned and internalized. That means individuals belonging to the same social aggregate are likely to acquire similar scripts and schemas and corresponding reactions.

Thus, when appraising social situations that have triggered scripts or schemas to become active, the appraisal process – from which emotions arise – is based on schematic processing. It operates on schemas instead of on "unbiased", raw data [52]. In such situations, it is likely that the resulting emotions do not reflect an individual's response to the actual "objective" person or situation, but rather the triggered schemas. This way, the amount to which social cognition is schematic and possibly erroneous may also affect emotional reactions and thus provoke "schematic emotions" [67].

Therefore, social cognition and socially distributed cognition lead to certain (classes of) emotions that do not reflect an "objective" appraisal of a person or situation, but instead the schemas of persons and the scripts of events an individual maintains. Because in social aggregates there is a high probability that many people share similar representations of scripts and schemas, also emotional reactions may bear features of these regularities. This way, social aggregates induce a certain amount of relatively homogenous emotional reactions to classes of acts, events, and objects.

What we have said in view of sociological models of emotion at the end of the preceding section also holds for the relation between social cognition and emotion. Though, we assume that the neurological dimension is more profound and stable, since the alteration of once established somatic-markers is hardly feasible. On the other hand, cognitive schemas, scripts, and consequent appraisals based thereon can be acquired and with greater effort also be altered throughout the lifespan. Therefore, they can serve as a means to adapt to fundamentally different social environments. Again, we will refer in more detail to the interrelation between sociological and cognitive theories in the following sections.

4.4 Social Control through Expression, Feedback, and Regulation

A component of the emotions we have not yet considered, although it is of utmost importance for the approach proposed here, is the communicative function of emotions. Until now, we have only dealt with cognitive and neural (that means internal) aspects of individuals' emotions and their consequences for social structures. But one of the most striking features of emotion is their communicative, i.e. *inter*individual function. We assume that the expression, communication, and regulation (coping) of emotion act as a crucial social control operator.

There is strong and consistent evidence that the expression of certain emotions such as anger, fear, enjoyment, sadness, and disgust – often called basic emotions – is distinctive and universal among the human species [20]. The expression of other emotions – sometimes called social emotions – such as shame, grief or embarrassment, does not seem to be universal among the human species, although patterns of expression are highly consistent in a cultural setting. However culture-specific expression of these emotions may have evolved, as long as individuals remain in the cultural setting they were socialized in, they can be almost certain to interpret emotion expressions in the appropriate way. Thus, emotions are a powerful communication device that signals to other individuals the emotional state an individual is in. Perceived emotion expressions allow with great certainty to infer a specific state of mind and the probable consequences for individual behavior, course of interaction, and group behavior.

Sociological and social psychological research conducted in the field of emotion expression has revealed several strategies actors use to deal with their emotions and emotions expressions. Hochschild [36] for example found out that*feeling rules* (or display rules), i.e. social norms, stipulate what an individual is supposed to feel in a specific interaction situation, and what emotions to display. Showing the appropriate emotions, that means the socially expected emotions in specific interaction situations, is mandatory for an individual in order to be socially accepted. Emotion work, or coping, is volitional cognitive effort to regulate and modulate both, the emotion actually felt and the facial and bodily display of an emotion, regardless whether the emotion on display is really felt or not.

The voluntary or involuntary display of an emotion is subject to social judgment by other individuals who perceive this emotion expression. Depending on what feeling rules are considered valid in a situation, an expression is judged to be either adequate or inadequate. Emotion expressions found to be inadequate signal that the individual expressing (and also probably experiencing) this emotion does not conform (mentally and behaviorally) to what is socially expected. Sanctions may be the consequence [21].

One possible sanctioning mechanism is, again, emotion. By showing anger for example, individuals can signal that they consider behavior to be deviant and not standard conforming. The result may be shame or embarrassment felt by the deviant individual. These emotions are supposed to encourage an individual to adapt its preceding behavior (emotion) by means of emotion work in order to be socially accepted.

The mechanisms illustrated show how emotions serve a reciprocal social control function: on the one hand as a norm-enforcement operator and sanctioning mechanism, on the other hand as an indicator that an individual's assessment and appraisal of a situation is not congruous with that of other individuals. By means of (emotional) sanctions and feedback an individual is then enforced to comply or to terminate an interaction. Social norms and feeling rules, being qualities of a social aggregate, therefore promote behavior regulation via emotions in order to maintain the qualities of a social aggregate.

This emotional control function acts on top of the mechanisms described before. The main difference compared to these mechanisms presumably is the degree to which emotional control is exerted and experienced consciously. Because of the interactive and immediate nature of this form of control, arousal is usually high and actors are aware of their (not necessarily volitional) emotional reactions.

The function of social control has already been described by [21], although in connection with his general social theory and not in view of an explicit sociological theory of emotion [68].

5 Emotion and the SAM Architecture

This section illustrates how the social components of emotion described in our integrative approach can be theoretically and conceptually applied to the multi-agent system architecture SAM (Socionic[1] Agent Model) [46]. We first give a very brief overview of the architectural modeling approach and then relate theoretical findings to the social theories that serve as a basis for the architecture.

[1] See [54] for an introduction to Socionics.

5.1 The Micro-Macro Link and the SAM Architecture

The SAM architecture is modeled by means of Petri nets using the "Renew" tool that allows direct implementation and execution of the model [49]. Here, reference nets (see [48] for a complete definition) – which are based on the "nets within nets"-paradigm as defined by [79] – are used to depict interdependencies between macro- and micro-level in hierarchical layers. We will focus on the three main social layers of the model that have been derived by an analysis of Bourdieu's [7] and Elias's [21] social theories: social structures, social processes, and actors. Originally, the intention was to use social theories to implement mechanisms of social control and habitual (i.e. organizational) behavior in order to analyze the interrelation between large-scale (macro-level) behavior of a multi-agent system and actions of individual agents (micro-level).

The ASKO (Behavior in Social Contexts) research group has modeled different aspects of these interrelations on an abstract conceptual level by describing social states, processes and acts [53]. Further investigations have shown that the sociological theories under examination provide an elaborated and extensive picture of large-scale processes. What is missing is how these features are represented inside individual actors. As long as one is concerned with modeling social structures, processes, acts and their interdependencies, this view is sufficient, but when it comes to modeling actual behavior generation or decision-making of individual agents, several problems arise.

Without a doubt, many valuable agent architectures already incorporate AI-based cognitive activities like planning, action-selection, emotion generation, etc. but unfortunately without making dedicated connections to macro-level phenomena [26]. Analysis of Elias's and Bourdieu's social theories has shown that several mechanisms they describe by which action and social structures are interlinked seem to have an "internal" functional counterpart in emotion. With our integrated approach to emotion presented above, the theories in question could be extended and given emotional foundations which also encompass cognitive and neural aspects. Results of the integration can be used to extend (and therefore probably to enhance) the multi-agent architecture and possibly also the sociological theories in a way that leads to an integration of emotional concepts and factors.

5.2 Habitus and Emotion

Central to the work of Bourdieu is the habitus-field theory with which he addresses the micro-macro link problem [7]. According to Bourdieu, the relationship between habitus and the logic of practice is crucial to understand micro-macro dynamics. The habitus is a cultural and social habitat that becomes internalized in the form of dispositions to act and behave, to think, reason, perceive, and even to feel in a certain way. The habitus can be seen as a set of socially determined bodily and mental dispositions that lack representational content and therefore seldom come to conscious awareness. If this should indeed be the case (e.g. through a field change or a personal crisis), it is important to note that not the habitus *itself* is atomized into a set of mental representations such as beliefs, desires, or intentions, but rather an individual forms beliefs *about* the habitus (and this belief-formation is again based on habitualreasoning).

Where does the habitus come from, then? It can be seen as the incorporation and internalization of the *logic of practice*. The logic of practice is a property of the *social field* within which all human action takes place. Basically, social fields are arenas for the struggle for resources and characterized by vertical stratification. Social fields operate by various mechanisms and rules which, taken together, form the logic of practice. The logic of practice defines the "borders" of a social field by issuing explicit and specific rules.

Individuals who have incorporated the logic of practice of a specific field provide practical acceptance of the practical logic of this specific field and thereby reproduce this very logic via the habitus. This way, a social field controls the behavior of its individuals. Thus, the habitus stabilizes its field, i.e. the field that originally produced the habitus [47].

This far Bourdieu's habitus-field theory has been modeled within the ASKO project [34]. As can be seen from the brief summary, the micro-macro dynamics described by the model resemble the micro-macro dynamics illustrated in our integrated approach to emotion. In view of general habitual behavior primarily the cognitive and neural components of emotion that we described seem to be relevant, whereas in view of the logic of practice and the social field, the regulation and control of emotion through norms and feedback deserve special attention.

We assume that the integrated approach to emotion presented here can serve as a neurocognitive foundation for some aspects of the habitus-field theory. Since the habitus is a phenomenon unconsciously guiding human behavior, we refer to our explanations of Damasio's somatic-marker hypothesis and the role of schemas, scripts, and schematic information processing. By interlinking both, general social theories and interdisciplinary research on emotion, a better understanding of micro-macro dynamics is achievable. The advantage is that this understanding is based on experimental and empirical evidence, rather than on theorizing alone.

5.3 Social Control and Emotion

One central aspect of Elias's [21] grand theory is the exertion of social control through norms and emotions. Tightly interlinked with social control and emotions is the reproduction and maintenance of social norms. According to Elias, any coherent social group (social aggregate) is characterized by struggles for status, power, prestige, social success, and appreciation. This rivalry leads to anxiety about the possible loss of one of these social resources. Elias assumes that this form of anxiety is inherent to the human species and goes back to attachment behavior in mother-infant relationships. Anxiety drives individuals in a social aggregate to constantly monitor other individuals' behavior in order to estimate one's own position in the social order relative to those of others. Knowledge of the positions of other individuals gives rise to efforts to maintain or even improve one's own position.

Crucial for an actor's position in the social order is willingness to comply with the norms of a social aggregate. Deviant behavior will be sanctioned by other individuals in various ways. On the one hand, the loss of social resources such as status, appreciation, and prestige may be the consequence. This may lead to negative emotions such as fear, anger or sadness. On the other hand, other individuals will also show negative emotions to express their discomfort with the deviant individual. Both, loss of resources and the expression of negative emotion may again have emotional consequences for the deviant individual: shame and embarrassment are the main emotions by which – according to Elias – social control is exerted.

Control then results in social bondage, i.e. "mental" bonds are created that tie an individual to the setting and configuration of a specific social aggregate. Fear of loosing group sympathy and support may transform the social bondage into a self-bondage, i.e. a volitional behavior regulation to be in accordance with prevailing social norms. This way, norms are exerted and exertion leads to the reproduction of a social norm [47,69].

Hinck and associates [35] have modeled this process of norm reproduction by means of high-level Petri nets. Further efforts towards modeling and incorporation of social norms and emotion in interface agents have been done by [57,59]. That work clearly shows the necessity to consider social norms for emotional expressive, socially intelligent agents in human-agent interactions. There, feeling rules are specified and related to an application, and requirements for agent- and user-modeling are outlined [58]. Staller and Petta [75] also have introduced emotions to the computational study of social norms, but from a slightly different perspective.

What has not been done yet is to further examine the role of emotions *per se* as a general indicator for deviant (internal or external) behavior. In the approach mentioned above, deviance is defined as overt behavior that clearly violates specific norms valid in a social aggregate. But when one relates Elias's theoretical findings to our integrated approach, it becomes obvious that the mere and possibly subtle display of an emotion in an interaction situation may indicate that an individual's assessment of a situation in general is not congruous with that of other individuals. This means, that for an actor to realize that another individual has assessed a situation differently from common social expectance, it is probably sufficient to perceive and interpret that individual's emotion expression at a certain time – obvious deviant external and norm-violating behavior is not necessary [37].

We assume that, according to appraisal theoretic approaches, emotions reflect an individual's perception and judgment of a social situation. In coherent groups, as explained, individuals constantly monitor each other's behavior to ensure norm compliance and to prepare eventual sanctions. Emotions are an early indicator that overt deviant actions might be carried out that could disrupt group coherence. They therefore allow interception and regulation at a stage where consequent and probably malicious actions have not yet been carried out.

Therefore, expression, perception, and judgment of emotions act as a control structure *on top* of neurocognitive components and their relation with the structure of a social aggregate. Emotional feedback, sanctions, feeling rules, and social norms make explicit and consciously available what has been impinged upon individuals in infancy and socialization. Social norms being one of the most important components of a social aggregate are tightly interlinked with emotions and are also reproduced via emotions. We thus conclude that the display of emotions and the feedback they provoke are also a vital component of the micro-macro link.

6 Conclusion and Outlook

We have presented an integrative approach to emotion that specifically aims at explaining the role of emotion in the micro-macro link. The micro-macro link is an unresolved key issue in the social sciences as well as in DAI research and addresses the problem of the relationship between individual action and social structure. Basically, the question is how regulation of individual behavior is achieved in a way that leads to social structural configurations allowing for phenomena such as cooperation, coherence or coordination, to name just some.

There are prominent theories in the social sciences that explicitly address this problem and which have been (partly) adopted by the DAI community (e.g. [28]). Even so it remains largely unsolved. With our approach to emotion that wedges the neural and cognitive underpinnings of the emotion process with sociological theories of emotion and general social theory, we contribute to an understanding of the micro-macro link that is most valuable for computer science since it draws on concepts that have been on the AI research agenda for quite a time.

Cognitive architectures and lately also emotional agents based on neurocognitive theories of emotion are drawing more and more attention. The work done in these areas and the problems of large-scale distributed multi-agent systems could have a stimulating effect on each other that has not been examined thoroughly. The problem still is that today there is neither a *theory* of emotion in sociology nor in the cognitive sciences that incorporates the diverse micro- and macro aspects and that could be used to design improved computational systems. Progress is made rapidly, as illustrated above, but mostly without consideration of the macro causes and consequences of emotion which are so important for DAI systems.

The core notion of our approach is that mental representations which are subject of cognitive activity and neural mechanisms and structures that enable as well as channel different modes of information processing in the brain are influenced by specific qualities of a social aggregate an individual is situated in. Priming, socialization, and social learning in various stages of the lifespan impinge the regularities found in a social aggregate, e.g. interaction chains, emotional reactions, judgments, stereotypes, norms, and rules, upon development dependent parts of the emotion process.

Characteristics and qualities of a social aggregate are emotionally represented in the way an individual's information processing architecture operates. Operation in this "biased" mode then works to maintain the structures and features that originally designed this mode of operation. There is no operation whatsoever free from social influence! Thus, social structures reproduce themselves via emotions and their foundational, (socially) formed and established neural and cognitive processes.

To specify the link between these mechanisms and dedicated grand theories in sociology, we have chosen Bourdieu and Elias as examples, since they have already described the micro-macro linkage with an emotional connotation. They obviously knew about the importance of the emotions but did not elaborate their role. In the ASKO project, parts of Bourdieu's habitus-field theory and Elias's social theory have been modeled. Here we improved the purely sociological interpretation and used emotion theories to relate psychological and neuroscientific work in such a way to these theories, that the micro-macro link gets a new and challenging perspective which might be adopted in the area of agent-oriented software engineering.

Further work has to be done to model the integrated approach to emotion in order to fit the existing models. But also for researchers using other methodologies, it is important to have a handy theory of emotion that can be used for any agent-oriented approach. The work presented here should be seen as a first step in this direction.

References

1. Aubé, M.; Senteni, A. (1996): Emotions as Commitments Operators: A Foundation for Control Structure in Multi-Agents Systems. In: van de Velde, W.; Perram, J.W. (Eds.): Agents Breaking Away. Proceedings of the 7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'96), January 22.-25., Eindhoven/NL. LNAI Vol. 1038. Berlin: Springer, 13–25.
2. Bainbridge, W.; Brent, E.; Carley, K.; Heise, D.; Macy, M.; Markovsky, B. and J. Skvoretz (1994): Artificial Social Intelligence. Annual Review of Sociology, 20: 407–436.
3. Barbalet, J.M. (1998): Emotion, Social Theory, and Social Structure. Cambridge: Cambridge University Press.
4. Bargh, J.A. (1997): The Automaticity of Everyday Life. In: Wyer, R.S. (Ed.): The Automaticity of Everyday Life. Advances in Social Cognition, Vol. X. Mahwah/NJ: Erlbaum, 1–62.
5. Bickmore, T., Picard, R. (2003): Subtle Expressivity by Relational Agents. In: Proceedings of the CHI 2003 Workshop on Subtle Expressivity for Characters and Robots. April 7., Fort Lauderdale, Florida/USA.
6. Bless, H. (2000): The Interplay of Affect and Cognition. The Mediating Role of General Knowledge Structures. In: Forgas, J.P. (Ed.): Feeling and Thinking. Cambridge: Cambridge University Press, 201–222.
7. Bourdieu, P. (1977): Outline of a Theory of Practice. Cambridge: Cambridge University Press.
8. Branscombe, N.R.; Cohen, B.M (1991): Motivation and Complexity Levels as Determinants of Heuristic Use in Social Judgment. In: Forgas, J.P. (Ed.): Emotion and Social Judgments. Oxford: Pergamon Press, 145–161.
9. Brothers, L. (1997): Friday's Footprint. New York: Oxford University Press.
10. Cacioppo, J.T.; Berntson, G.G.; Adolphs, R.; Carter, C.S.; Davidson, R.J.; McClintock, M.K.; McEwen, B.S.; Meaney, M.J.; Schacter, D.L.; Sternberg, E.M.; Suomi, S.S. and S.E. Taylor (Eds.)(2002): Foundations in Social Neuroscience. Cambridge/MA: The MIT Press.
11. Canamero, D.; van de Velde, W. (2000): Emotionally Grounded Social Interaction. In: Dautenhahn, K. (Ed.): Human Cognition and Social Agent Technology. Advances in Consciousness Research, Vol. 19. Amsterdam: John Benjamins, 137–162.
12. Castelfranchi, C.; Werner, E. (1994): The MAAMAW Spirit and this Book. In: Castelfranchi, C.; Werner, E. (Eds.): Artificial Social Systems. Selected Papers from the 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'92), July 29.–31., S. Martino al Cimino, Italy. LNAI Vol. 830. Berlin: Springer,VII–XVII.
13. Clark, M.S.; Brisette, I. (2000): Relationship beliefs and emotion: Reciprocal effects. In: Frijda, N.H.; Manstead, A.S.; Bem, S. (Eds.): Emotions and Beliefs. Cambridge: Cambridge University Press, 212–240.
14. Clore, G.L.; Schwarz, N.; Conway, M. (1994): Affective Causes and Consequences of Social Information Processing. In: Wyer, R.S.; Srull, T.K. (Eds.): Handbook of Social Cognition. 2nd Ed. Hillsdale/NJ: Lawrence Erlbaum, 323–417.
15. Collins, R. (1981): On the Microfoundations of Macrosociology. American Journal of Sociology, 86: 984–1014.
16. Collins, R. (1984): The Role of Emotion in Social Structure. In: Scherer, K.R.; Ekman, P. (Eds.): Approaches to Emotion. Hillsdale: Lawrence Erlbaum, 385–396.
17. Conte, R.; Castelfranchi, C. (1995): Norms as Mental Objects – From Normative Beliefs to Normative Goals. In: Castelfranchi, C.; Müller, J.-P. (Eds.): From Reaction to Cognition. Selected Papers from the 5th European Workshop on Modeling Autonomous Agents in a Multi-Agent World (MAAMAW'93), August 25.-27., Neuchatel, Switzerland. LNAI Vol. 957. Berlin: Springer, 186–196.

18. Damasio, A.R. (1994): Descartes' Error. New York: Grosset/Avon.
19. Dautenhahn, K. (2000): Socially Intelligent Agents and the Primate Social Brain – Towards a Science of Social Minds. In: Dautenhahn, K. (Ed.): Proceedings of the AAAI Fall Symposium "Socially Intelligent Agents - The Human in the Loop". Technical Report FS-00-04. Menlo Park: AAAI Press, 35–51.
20. Ekman, P. (1989): The argument and evidence about universals in facial expressions of emotion. In: Wagner, H.; Manstead, A.S. (Eds.): Handbook of Social Psychophysiology. Chichester: Wiley, 143–164.
21. Elias, N. (1978): The Civilizing Process. Oxford: Blackwell.
22. Ferber, J. (1999): Multi-Agent Systems. Harlow: Addison-Wesley.
23. Fiske, ST.; Taylor, S.E. (1991): Social Cognition. New York: McGraw-Hill.
24. Forgas, J.P. (1992): Affect in Social Judgements and Decisions: A Multi-Process Model. In: Zanna, M.P. (Ed.): Advances in Experimental Social Psychology. New York: Academic Press, 227–275.
25. Franks, D.D. (1999): Some Convergences and Divergences Between Neuroscience and Symbolic Interaction. In: Franks, D.D.; Smith, T.S. (Eds.): Mind, Brain, and Society: Toward a Neurosociology of Emotion. Social Perspectives on Emotion, Vol. 5. Stamford/CT: JAI Press, 157–182.
26. Gasser, L. (1991): Social Conceptions of Knowledge and Action: DAI Foundations and Open System Semantics. Artificial Intelligence, 47(1-3): 107–138.
27. Gazzaniga, M.S. (1985): The Social Brain. New York/NY: Basic Books.
28. Giddens, A. (1986): The Constitution of Society. Berkeley/CA: University of California Press.
29. Gmytrasiewicz, P.J.; Lisetti, C.L. (2000): Using Decision Theory to Formalize Emotions for Multi-Agent Systems. In: Proceedings of the 4th Intl. Conference on Multi-Agent Systems (ICMAS'00), July 10.–12., Boston/MA. IEEE Press.
30. Gordon, S.L. (1990): Social Structural Effects on Emotion. In: Kemper, T.D. (Ed.): Research Agendas in the Sociology of Emotions. Albany/NY: State University of New York Press, 145–179.
31. Haddadi, A.; Sundermeyer, K. (1996): Belief-Desire-Intention Agent Architectures. In: O'Hare, G.M.; Jennings, N.R. (Eds.): Foundations of Distributed Artificial Intelligence. New York: Wiley & Sons, 169–186.
32. Hammond, M. (1990): Affective Maximization. A New Macro-Theory in the Sociology of Emotion. In: Kemper, T.D. (Ed.): Research Agendas in the Sociology of Emotions. Albany/NY: State University of New York Press, 58–81.
33. Heise, D.R. (1979): Understanding Events. New York: Cambridge University Press.
34. Hinck, D.; Köhler, M.; Langer, R.; Moldt, D. and H. Rölke (2002): Modellierungen und Reanalysen zur Habitus-Feld Theorie von Pierre Bourdieu. Working Paper of the ASKO Research Group, University of Hamburg, Computer Science Department.
35. Hinck, D.; Köhler, M.; Langer, R.; Moldt, D. and H. Rölke (2001): Organisation etablierter Machtzentren: Modellierungen und Reanalysen zu Norbert Elias. Working Paper of the ASKO Research Group, University of Hamburg, Computer Science Department. FBI-HH-306/01.
36. Hochschild, A.R. (1979): Emotion Work, Feeling Rules, and Social Structure. American Journal of Sociology, 85(3): 551–575.
37. Horstmann, G. (2003): What Do Facial Expressions Convey: Feeling States, Behavioral Intentions, or Action Requests? Emotion, 3(2): 150–166.
38. Isen, A.M.; Means, B. (1983): The Influence of Positive Affect on Decision-Making Strategy. Social Cognition, 2: 18–31.
39. Johnson-Laird, P.N.; Oatley, K. (1992): Basic Emotions, Rationality, and Folk Theory. Cognition & Emotion, 6(3/4): 201–223.

40. Kaiser, S.; Wehrle, T. (2001): The Role of Facial Expression in Intra-Individual and Inter-Individual Emotion Regulation. In: Canamero, D. (Ed.): Emotional and Intelligent II: The Tangled Knot of Social Cognition. Papers from the 2001 AAAI Fall Symposium. Technical Report FS-01-02. Menlo Park: The AAAI Press, 61–66.
41. Keltner, D.; Haidt, J. (1999): Social Functions of Emotion at Four Levels of Analysis. Cognition & Emotion, 13(5): 505–521.
42. Kemper, T.D. (1978): A Social Interactional Theory of Emotions. New York: Wiley.
43. Kemper, T.D. (1981): Social Constructionist and Positivist Approaches to the Sociology of Emotions. American Journal of Sociology, 87(2): 336–362.
44. Kleining, G. (1994): Qualitativ-heuristische Sozialforschung: Schriften zur Theorie und Praxis. Hamburg: Fechner.
45. Knorr-Cetina, K.D. (1981): Introduction: The micro-sociological challenge of macro-sociology. In: Knorr-Cetina, K.D.; Cicourel, A.V. (Eds.): Advances in social theory and methodology. Toward an integration of micro- and macro-sociologies. Boston/MA: Routledge & Kegan Paul, 1–47.
46. Köhler, M.; Moldt, D.; Rölke, H. (2001): Modeling the Structure and Behaviour of Petri Net Agents. In: Colom, J.M.; Maciej, K. (Eds.): Proceedings of the 22nd International Conference on Application and Theory of Petri Nets (ICATPN'01), June 25.-29., Newcastle/UK. LNCS Vol. 2075. Berlin: Springer, 224–241.
47. Köhler, M.; Rölke, H. (2002): Modelling the micro-macro-link: Towards a sociologically grounded design of multi agent systems. In: Jonker, C; Lindemann, G.; Panzarasa, P. (Eds.): Proceedings of the Workshop Modelling Artificial Societies and Hybrid Organizations (MASHO'02), held in conjunction with the 25th German Conference on Artificial Intelligence, 2002.
48. Kummer, O. (2002): Referenznetze. Berlin: Logos-Verlag.
49. Kummer, O.; Wienberg, F. (1998): Reference net workshop (RENEW). University of Hamburg, Computer Science Department. http://www.renew.de
50. Lazarus, R.S. (1984): Thoughts on the Relations Between Emotion and Cognition. In: Scherer, K.R.; Ekman, P. (Eds.): Approaches to Emotion. Hillsdale: Lawrence Erlbaum, 247–257.
51. LeDoux, J. (1996): The Emotional Brain. New York: Simon & Schuster.
52. Leventhal, H.; Scherer, K.R. (1987): The Relationship of Emotion to Cognition. A Functional Approach to a Semantic Controversy. Cognition & Emotion, 1(1): 3–28.
53. Lüde, R. von; Moldt, D.; Valk, R. (Eds.)(2003): Sozionik – Modellierung soziologischer Theorie. Münster: LIT-Verlag.
54. Malsch, T. (2001): Naming the Unnameable: Socionics and the Sociological of/to Distributed Artificial Intelligence. Autonomous Agents and Multi-Agent Systems, 4(3): 155–186.
55. Miller, R.S.; Leary, M.R. (1992): Social Sources and Interactive Functions of Emotion. The Case of Embarrassment. In: Clark, M.S. (Ed.): Emotion and Social Behavior. Review of Personality and Social Psychology, 14. Newbury Park/CA: Sage, 202–221.
56. Minsky, M. (1986): The Society of Mind. New York: Simon & Schuster.
57. Moldt, D.; von Scheve, C. (2001): Emotional Actions for Emotional Agents. In: Agents & Cognition. Proceedings of the AISB'01 Symposium on Emotion, Cognition, and Affective Computing. York/UK: SSAISB Press, 121–128.
58. Moldt, D.; von Scheve, C. (2001a): Emotions and Multimodal Interface-Agents: A Sociological View. In: Oberquelle, H.; Oppermann, R.; Krause, J. (Eds.): Mensch & Computer 2001. Proceedinsg of the 1st Interdisciplinary Conference. Stuttgart: Teubner, 287–295.

59. Moldt, D.; von Scheve, C. (2002): Attribution and Adaptation: The Case of Social Norms and Emotion in Human-Agent Interaction. In: Marsh, S.; Meech, J.F.; Nowell, L. and K. Dautenhahn (Eds.): Proceedings of The Philosophy and Design of Socially Adept Technologies, workshop held in conjunction with CHI'02, April 4th, Minneapolis/Minnesota, USA. National Research Council Canada, NRC #44918, pp 39–41.
60. Moldt, D.; von Scheve, C. (2002a): Emotions in Hybrid Social Aggregates. In: Herczeg, M.; Prinz, W.; Oberquelle, H. (Eds.): Mensch & Computer 2002. Vom interaktiven Werkzeug zu kooperativen Arbeits- und Lernwelten. Stuttgart: Teubner, 343–352.
61. Nesse, R.M. (1994): Computer Emotions and Mental Software. Social Neuroscience Bulletin, 7(2): 36–37.
62. Oatley, K. (2000): The sentiments and beliefs of distributed cognition. In: Frijda, N.H.; Manstead, A.S.; Bem, S. (Eds.): Emotions and Beliefs. Cambridge: Cambridge University Press, 78–107.
63. Roseman, I.J.; Smith, C. A. (2001): Appraisal theory: Overview, assumptions, varieties, controversies. In: Scherer, K.R.; Schorr, A.; Johnstone, T. (Eds.): Appraisal Processes in Emotion: Theory, Methods, Research. Oxford: Oxford University Press, 3–19.
64. Parsons, T. (1951): The Social System. New York: Free Press.
65. Prendinger, H.; Ishizuka, M. (2002): Evolving social relationships with animate characters. In: Proceedings of the Symposium of the AISB'02 Convention on Animating Expressive Characters for Social Interactions, London/UK, 2002, 73–78.
66. Picard, R.W. (1997): Affective Computing. Cambridge/MA: The MIT Press.
67. Reisenzein, R. (2001): Appraisal processes conceptualized from a schema theoretic perspective: Contributions to a process analysis of emotions. In: Scherer, K.R.; Schorr, A.; Johnstone, T. (Eds.): Appraisal Processes in Emotion: Theory, Methods, Research. Oxford: Oxford University Press, 187–204.
68. Scheff, T.J. (1988): Shame and Conformity: The Deference-Emotion System. American Sociological Review, 53: 395–406.
69. Scheff, T.J. (1997): Emotions, the Social Bond, and Human Reality. Cambridge: Cambridge University Press
70. Scheutz, M.; Logan, B. (2001): Affective vs. Deliberative Agent Control. In: Agents & Cognition. Proceedings of the AISB'01 Symposium on Emotion, Cognition, and Affective Computing. York/UK: SSAISB Press, 1–10.
71. Schillo, M.; Fischer, K.; Klein C. (2001): The Micro-Macro Link in DAI and Sociology. In: Moss, S.; Davidsson, P. (Eds.): Multi-Agent Based Simulation. Second International Workshop on Multi-Agent Based Simulation, July 2000, Boston/USA. LNAI Vol. 1979. Berlin: Springer, 133–148.
72. Shott, S. (1979): Emotion and Social Life: A Symbolic Interactionist Analysis. American Journal of Sociology, 84(4): 1317–1334.
73. Sloman, A.; Logan, B. (2000): Evolvable Architectures for Human-like Minds. In: Hatano, G.; Okada, N.; Tanabe, H. (Eds.): Affective Minds. Proceedings of the 13th Toyota Conference. Amsterdam: Elsevier, 169–181.
74. Smith, C.A.; Pope, L.K. (1992): Appraisal and Emotion. The Interactional Contributions of Dispositional and Situational Factors. In: Clark, M.S. (Ed.): Emotion and Social Behavior. Review of Personality and Social Psychology, 14. Newbury Park/CA: Sage, 32–62.
75. Staller A., Petta P. (2001): Introducing Emotions into the Computational Study of Social Norms: A First Evaluation. Journal of Artificial Societies and Social Simulation, 4(1). http://jasss.soc.surrey.ac.uk/4/1/2.html
76. Tredway, J.V.; Knapp, S.J.; Tredway, L.C. and D.L. Thomas (1999): The Neurosociological Role of Emotions in Early Socialization, Reason, Ethics, and Morality. In: Franks, D.D.; Smith, T.S. (Eds.): Mind, Brain, and Society: Toward a Neurosociology of Emotion. Social Perspectives on Emotion, Vol. 5. Stamford/CT: JAI Press, 109–156.
77. Turner, J.H. (1988): A Theory of Social Interaction. Stanford: Stanford University Press.

78. Turner, J.H. (1999): Toward a General Sociological Theory of Emotions. Journal for the Theory of Social Behavior, 29(2): 133–161.
79. Valk, R. (1998): Petri nets as token objects: An introduction to elementary object nets. In: Desel, J.; Silva, M. (Eds.): Application and Theory of Petri nets. LNCS Vol. 1420. Berlin: Springer, 1–25.
80. Ventura, R.; Pinto-Ferreira, C. (1999): Emotion-Based Agents: Three Approaches to Implementation. In: Proceedings of the Workshop on Emotion-Based Agent Architectures (EBAA'99), held in conjunction with Autonomous Agents (Agents'99), 1st May, Seattle/USA.
81. Weber, M. (1976): Wirtschaft und Gesellschaft. Grundriß der verstehenden Soziologie. 5th. revised edition. Tübingen: Mohr.
82. Wooldridge, M. (2000): Reasoning about Rational Agents. Cambridge/MA: The MIT Press.
83. Zajonc, R. (1984): On Primacy of Affect. In: Scherer, K.R.; Ekman, P. (Eds.): Approaches to Emotion. Hillsdale/NJ: Lawrence Erlbaum, 259–270

What Is a Normative Goal?
Towards Goal-Based Normative Agent Architectures

Mehdi Dastani[1] and Leendert van der Torre[2]

[1] Institute of Information and Computer Sciences, Utrecht University
mehdi@cs.uu.nl
[2] CWI Amsterdam
torre@cwi.nl

Abstract. In this paper we are interested in developing goal-based normative agent architectures. We ask ourselves the question what a normative goal is. To answer this question we introduce a qualitative normative decision theory based on belief (B) and obligation (O) rules. We show that every agent which makes optimal decisions – which we call a BO rational agent – acts *as if* it is maximizing the set of normative goals that will be achieved. This is the basis of our design of goal-based normative agents.

1 Introduction

Simon [32] interpreted goals as utility aspiration levels, in planning goals have a notion of desirability as well as intentionality [20], and in the Belief-Desire-Intention or BDI approach [11,29] goals have been identified with desires. Moreover, recently several approaches have been introduced to extend decision making and planning with goal generation [14]. For example, Thomason's BDP logic [33] extends the BDI approach with goal generation and planning, and Broersen *et al.* 's BOID architecture [9] elaborates on the goal generation mechanism for a more general class of cognitive agents. But what is this thing called goal? Although there are many uses of goals in planning and more recently in agent theory, the ontological status of goals seems to have received little attention.

In this paper we try to find out what a normative goal is by comparing normative decision systems with knowledge-based systems in which decisions are considered to be the result of goal based planning. Of course, such a comparison is complicated by the fact that there are many different kinds of normative and knowledge-based systems. We therefore restrict ourselves to the characterizations illustrated in Figure 1.

This figure should be read as follows. First, for our comparison normative agents are decision-making agents in normative systems which perform practical reasoning [34,37]. They can formally be described by a reasoning mechanism based on (defeasible) deontic logic which describes the relation between a set of beliefs *(B)* including observations, a set of obligations *(O),* and decisions or actions. If we replace the set of obligations by a set of desires *(D),* then many qualitative decision theories such as [25,33,36] also fit this description. Second, knowledge-based agents have as input a knowledge base *(KB)* including observations and goals, and they have as output actions or plans.

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 210–227, 2004.

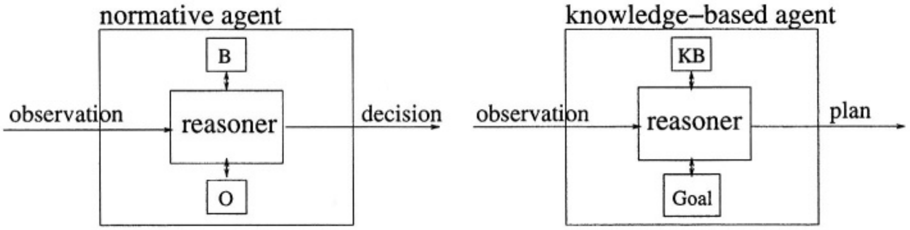


Fig. 1. Agent

Knowledge-based systems have been advocated by Newell and Simon [28,32] and have been implemented by for example SOAR [23] and ACT [1]. Moreover, more recent BDI systems like PRS [21] fit this description. The main reasoning task of the knowledge-based system is planning.

How can we compare these two kinds of decision making systems? First we unify the beliefs with the knowledge base, because both represent the motivational attitude of the system. Moreover, we unify decisions with actions and plans. The main problem is the unification of the motivational attitude, the obligations (or, in other qualitative decision theories, the desires) and the goals. Rao and Georgeff [29] propose, at a very high level of abstraction, that desires and goals can be unified. However, this has been criticized by several authors [16,19,22]. An argument against the unification is that desires can conflict whereas goals in Rao and Georgeff's framework cannot. Another argument is that goals may be adopted from another agent, whereas desires cannot be adopted. Moreover, desires are more stable than goals [13].

Thomason [33] proposes a logical theory in which desires are a more primitive concept than goals, in the sense that goals can be inferred from desires. Broersen *et al.* [9] extend this argument to obligations and propose an architecture in which goals can be inferred from desires, intentions and obligations and in which goal generation gets a prominent place. For our comparison, we define goal generation as a theory with input beliefs, observations and obligations, and as output goals. Now we can use the output of goal generation as input for the knowledge-based system to infer decisions or actions. The idea can be paraphrased by:

Goal-based decision making is goal generation together with goal-based planning

This decomposition of decision making in goal generation and planning raises several questions, such as:

- How to represent beliefs? How to represent obligations? In this paper we represent beliefs and obligations by rules, following the dominant tradition in deontic logic (see e.g. [26,27]).
- How to develop a normative decision theory based on belief and obligation rules? In this paper we introduce a qualitative decision theory, based on belief (B) and obligation (O) rules.

- How can this decision theory be decomposed into goal generation and goal-based planning? How to define a notion of normative goals in this theory? In this paper, we show how these questions can be answered for our qualitative decision theory.

Our main aim in this paper is not to convince the reader that this decision theory is the best option available. It has the advantage that it is a simple theory, but it is not the most advanced one available in the literature. Our aim is to show how, given a decision theory, a distinction can be made between goal generation and goal-based planning. The motivation of our study is to give formal foundations for goal-based normative agent architectures, such as the one depicted in Figure 2.

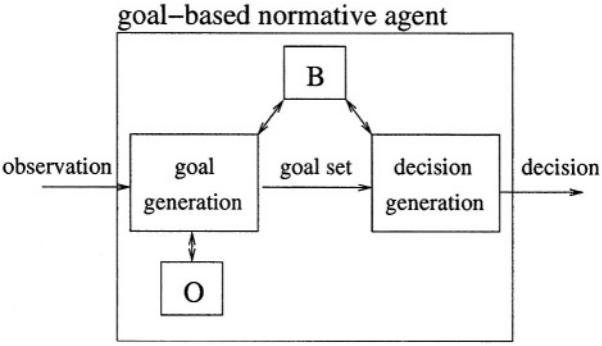


Fig. 2. Goal-based agent

This figure should be read as follows. The input of the system is an observation and its output is a decision (or action, or plan). There are two components, which we call goal generation and decision generation. Goal generation has a goal set as its output, which is the input for decision generation. Decision generation is for example the reasoner or the planner of the classic knowledge-based system depicted in Figure 1. Decision making is based on two sets of rules, represented by components B for belief rules and O for obligation rules. In particular, both goal generation and decision generation use belief rules, but only goal generation uses obligation rules. This represents that the motivational attitude encoded in O is transformed by goal generation in the goal set. In this paper, the difference between obligation rules and normative goal set is that sets of obligation rules are sets of pairs of propositional formulas, whereas a goal set is a set of propositional formulas, or, when we distinguish positive and negative goals, two sets of propositional sentences.

Like classical decision theory, but in contrast to several proposals in the BDI approach [11,29], the theory does not incorporate temporal reasoning and scheduling.

The layout of this paper is as follows. We first develop a normative logic of decision. This logic tells us what the optimal decision is, but it does not tell us how to find this optimal decision. We then consider the AI solution to this problem [32]: break down the decision problem into goal generation and goal-based decisions.

2 A Normative Decision Theory

The qualitative decision theory introduced in this section is based on sets of belief and obligation rules. There are several choices to be made, where our guide is to choose the simplest option available.

2.1 Decision Specification

The starting point of any theory of decision is a distinction between choices made by the decision maker (flip a coin) and choices imposed on it by its environment (head or tail). We therefore assume the two disjoint sets of propositional atoms $A = \{a, b, c, \ldots\}$ (the agent's decision variables [24] or controllable propositions [8]) and $W = \{p, q, r, \ldots\}$ (the world parameters or uncontrollable propositions). We write:

- L_A, L_W and L_{AW} for the propositional languages built up from these atoms in the usual way, and $x, y, \ldots$ for any sentences of these languages.
- Cn_A, Cn_W and Cn_{AW} for the consequence sets, and $\models_A$, $\models_W$ and $\models_{AW}$ for satisfiability, in any of these propositional logics.
- $x \Rightarrow y$ for an ordered pair of propositional sentences called a rule.

A decision specification given in Definition 1 is a description of a decision problem. It contains a set of belief and obligation rules, as well as a set of facts and an initial decision (or prior intentions). A belief rule 'the agent believes y in context x' is an ordered pair $x \Rightarrow y$ with $x \in L_{AW}$ and $y \in L_W$, and an obligation rule 'the agent ought y in context x' is an ordered pair $x \Rightarrow y$ with $x \in L_{AW}$ and $y \in L_{AW}$. It implies that the agent's beliefs are about the world ($x \Rightarrow p$), and not about the agent's decisions. These beliefs can be about the effects of decisions made by the agent ($a \Rightarrow p$) as well as beliefs about the effects of parameters set by the world ($p \Rightarrow q$). Moreover, the agent's obligations can be about the world ($x \Rightarrow p$, obligation-to-be), but also about the agent's decisions ($x \Rightarrow a$, obligation-to-do). These obligations can be triggered by parameters set by the world ($p \Rightarrow y$) as well as by decisions made by the agent ($a \Rightarrow y$).

The reason that we do exclude decision variables in the consequent of the belief rules is that belief rules are assumed here to be not defeasible: a belief for decision a cannot be defeated by the decision $\neg a$. This condition can be relaxed in an extension of the theory which incorporates defeasible belief rules.

Definition 1 (Decision specification). *A decision specification is a tuple $DS = \langle F, B, O, d_0 \rangle$ that contains a consistent set of facts $F \subseteq L_W$, a finite set of belief rules $B \subseteq L_{AW} \times L_W$, a finite set of obligation rules $O \subseteq L_{AW} \times L_{AW}$ and an initial decision $d_0 \subseteq L_A$.*

2.2 Decisions

The belief rules are used to express the expected consequences of a decision, where a decision d is any subset of L_A that implies the initial decision d_0, and the set of expected consequences of this decision d is the belief extension of $F \cup d$, as defined in Definition 2 below. Belief rules are interpreted as inference rules. We write $E_R(S)$ for the R extension of S.

Definition 2 (Extension). *Let $R \subseteq L_{AW} \times L_{AW}$ be a set ofrules andS $\subseteq L_{AW}$ a set of sentences. The consequents of the S-applicable rules are:*

$$R(S) = \{y \mid x \Rightarrow y \in R, x \in S\}$$

and the R extension of S is the set of the consequents of the iteratively S-applicable rules:

$$E_R(S) = \cap_{S \subseteq X, R(Cn_{AW}(X)) \subseteq X} X$$

The following proposition shows that $E_R(S)$ is the smallest superset of *S* closed under the rules *R* interpreted as inference rules.

Proposition 1 (Iteration). *Let*

- $E_R^0(S) = S$
- $E_R^i(S) = E_R^{i-1}(S) \cup R(Cn_{AW}(E_R^{i-1}(S)))$*for* $i > 0$

We have $E_R(S) = \cup_0^{\infty} E_R^i(S)$.

Proof. *Followsfrom analogous results in input/output logic [26].*

The following proposition shows that $E_R(S)$ ismonotonic.

Proposition 2 (Monotonicity). *We have* $R(S) \subseteq R(S \cup T)$ *and* $E_R(S) \subseteq E_R(S \cup T)$.

Proof. *Follows directlyfrom the definition.*

Monotonicity is illustrated by the following example.

Example 1. Let $A = \{a\}$, $W = \{p\}$, $R = \{\top \Rightarrow p, a \Rightarrow \neg p\}$ and $S = \{a\}$, where $\top$ stands for any tautology like $p \vee \neg p$. We have $E_R(S) = \{a, p, \neg p\}$, thus the *R* extension of *S* is inconsistent. We do *not* have that for example the specific rule overrides the more general one such that $E_R(S) = \{a, \neg p\}$.

We assume that a decision is an arbitrary subset of controllable propositions that implies the initial decision and does not imply a contradiction in its belief consequences.

Definition 3 (Decisions). *Let* $DS = \langle F, B, O, d_0 \rangle$ *be a decision specification. The set of DS decisions is*

$$\Delta = \{d \mid d_0 \subseteq d \subseteq L_A, E_B(F \cup d) \textbf{\textit{ is consistent}} \}$$

When a decision implies a, then we say that the agent makes decision a, or that it does a. The following example illustrates decisions.

Example 2. Let $A = \{a, b, c\}$, $W = \{p, q, r\}$ and $DS = \langle F, B, O, d_0 \rangle$ with $F = \{p \rightarrow r\}$, $B = \{b \Rightarrow \neg q, c \Rightarrow p, p \Rightarrow q\}$, $O = \{\top \Rightarrow a, a \Rightarrow b, \top \Rightarrow r\}$ and $d_0 = \{a\}$. The initial decision d_0 reflects that the agent has already decided in an earlier stage to reach the obligation $\top \Rightarrow a$. Note that the consequents of all *B* rules are sentences of L_W, whereas the antecedents of the *B* rules as well as the antecedents and consequents of the *O* rules are sentences of L_{AW}. We have due to the definition of $E_R(S)$:

$E_B(F \cup \{a\}) = \{p \rightarrow r, a\}$
$E_B(F \cup \{a, b\}) = \{p \rightarrow r, a, b, \neg q\}$
$E_B(F \cup \{a, c\}) = \{p \rightarrow r, a, c, p, q\}$
$E_B(F \cup \{a, b, c\}) = \{p \rightarrow r, a, b, c, p, q, \neg q\}$

Note that $\{a, b, c\}$ is not a *DS* decision, because its extension is inconsistent.

2.3 Optimal Decisions

Given the specification of a decision problem, Definition 3 indicates all possible decisions that can be generated. In the following, we introduce a normative decision theory, which determines the interpretation of the elements of the decision specification. This normative decision theory imposes an ordering on possible decisions based on the obligation rules and provides a way to identify optimal decisions. In particular, the obligation rules are used to compare the decisions. There are various ways to compare decisions based on the obligation rules. For example, one can compare decisions by considering the obligation rules that are violated by them where an obligation rule $x \Rightarrow y$ is called to be violated by a decision if the belief consequences of the decision imply $x \wedge \neg y$. Another way to compare decisions is by considering the reached obligation rules where an obligation rule $x \Rightarrow y$ is called to be reached by a decision if the belief consequences of the decision imply $x \wedge y$. In this paper, we compare decisions by considering the unreached obligation rules. An obligation rule $x \Rightarrow y$ is unreached by a decision if the belief consequences of the decision imply x, but not y. Note that the set of unreached desires is a superset of the set of violated desires.

Definition 4 (Comparing decisions). *Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification and d be a DS decision. The unreached obligations of decision d are:*

$$U(d) = \{x \Rightarrow y \in O \mid E_B(F \cup d) \models x \textit{ and } E_B(F \cup d) \not\models y\}$$

Decision d_1 is at least as good as decision d_2, written as $d_1 \geq_U d_2$, iff

$$U(d_1) \subseteq U(d_2)$$

Decision d_1 dominates decision d_2, written as $d_1 >_U d_2$, iff

$$d_1 \geq_U d_2 \textit{ and } d_2 \not\geq_U d_1$$

Decision d_1 is as good as decision d_2, written as $d_1 \sim_U d_2$, iff

$$d_1 \geq_U d_2 \textit{ and } d_2 \geq_U d_1$$

The following continuation of Example 2 illustrates the comparison of decisions.

Example 3 (Continued). We have:
$U(\{a\}) = \{\top \Rightarrow r, a \Rightarrow b\}$,
$U(\{a,b\}) = \{\top \Rightarrow r\}$,
$U(\{a,c\}) = \{a \Rightarrow b\}$.
We thus have that the decisions $\{a,b\}$ and $\{a,c\}$ both dominate the initial decision $\{a\}$, i.e. $\{a,b\} >_U \{a\}$ and $\{a,c\} >_U \{a\}$, but the decisions $\{a,b\}$ and $\{a,c\}$ do not dominate each other nor are they as good as each other, i.e. $\{a,b\} \not\geq_U \{a,c\}$ and $\{a,c\} \not\geq_U \{a,b\}$.

The following proposition shows that the binary relation $\geq_U$ on decisions is transitive and we can thus interpret it as a preference relation.

Proposition 3 (Transitivity). *The binary relation $\geq_U$ is transitive.*

Proof. *Follows from transitivity of subset-relation.*

A consequence of this normative decision theory is that the ordering of decisions is influenced only by the subset of obligation rules which is disjoint with the set of belief rules. The following proposition shows that obligations only matter as long as they are different from beliefs.

Proposition 4 (Redundancy). *Let DS $= \langle F, B, O, d_0\rangle$ and $DS' = \langle F, B, O \setminus B, d_0\rangle$. Then, d is a DS decision iff d is a DS' decision. Moreover, for two DS decisions d_1 and d_2, $d_1 \geq_U d_2$ in DS iff $d_1 \geq_U d_2$ in DS'.*

Proof. *By Definition 3 DS and DS' have the same set of decisions. Let $x \Rightarrow y \in B$ and $x \Rightarrow y \in O$ in both DS and DS'. Then, Proposition 1 states that if $x \in E_B(d \cup F)$ then also $y \in E_B(d \cup F)$. Consequently, for DS and DS' the rule $x \Rightarrow y$ cannot be in $U(d)$ and thus this rule cannot change the ordering relation $\geq_U$ in DS or DS'.*

The decision theory prescribes an economic rational decision maker to select the optimal or best decision, which is defined as a decision that is not dominated.

Definition 5 (Optimal decision). *Let DS be a decision specification. A DS decision d is U-optimal iff there is no DS decision d' with $d' >_U d$.*

The following example illustrates optimal decisions.

Example 4. Let $A = \{a, b\}$, $W = \emptyset$ and $DS = \langle \emptyset, \emptyset, \{a \Rightarrow b\}, \emptyset\rangle$. We have that $U(d) = \{a \Rightarrow b\}$ if $d \models_{AW} a$ and $d \not\models_{AW} b$, $U(d) = \emptyset$ otherwise. The U-optimal decisions are the decisions d that either do not imply a or that imply $a \wedge b$.

The following proposition shows that for each decision specification, there is at least one optimal decision. This is important, because it guarantees that agents can always act in some way.

Proposition 5 (Existence). *Let DS be a decision specification. There is at least one U-optimal DS decision.*

Proof. *Since the facts F are consistent, there exists at least one DS decision. Since the set of desire rules is finite, there do not exist infinite ascending chains in $\geq_U$, and thus there is an U-optimal decision.*

For a given decision specification, there may be more than one optimal decision. Therefore, we introduce an alternative to our notion of optimality by adding minimality in the definition of optimal decisions. Definition 6 introduces a distinction between smaller and larger decisions. A smaller decision implies that the agent commits itself to less choices. A minimal optimal decision is an optimal decision such that there is no smaller optimal decision.

Definition 6 (Minimal optimal decision). *A decision d is a minimal U-optimal DS decision iff it is an U-optimal DS decision and there is no U-optimal DS decision d' such that $d \models d'$ and $d' \not\models d$.*

The following example illustrates the distinction between optimal and minimal optimal decisions.

Example 5. Let $A = \{a, b\}$, $W = \{p\}$ and $DS = \langle F, B, O, d_0 \rangle$ with $F = \emptyset$, $B = \{a \Rightarrow p, b \Rightarrow p\}$, $O = \{\top \Rightarrow p\}$, $d_0 = \emptyset$. Optimal decisions are $\{a\}$, $\{b\}$ and $\{a, b\}$, of which only the former two are minimal. Note that $\{a \vee b\}$ is not an optimal decision, because $p \not\in E_B(\{a \vee b\})$.

The following proposition illustrates in what sense a decision theory based on optimal decisions and one based on minimal optimal decisions are different.

Proposition 6 (Minimality). *There is a decision specification DS with an U-optimal DS decision* d*, such that there is no minimal U-optimal DS decision* d' *with* $d \sim_U d'$.

Proof. Consider $DS = \langle \emptyset, \emptyset, \{\top \Rightarrow a, a \Rightarrow \neg a\}, \emptyset \rangle$. *The unique minimal U-optimal decision is* $d_1 = \emptyset$. *The decision* $d_2 = \{a\}$ *is also U-optimal, but we do not have* $d \sim_U d'$.

The following example illustrates that the minimal decision d_0 is not necessarily an optimal decision.

Example 6. Let $A = \{a\}$, $W = \emptyset$ and $DS = \langle \emptyset, \emptyset, \{\top \Rightarrow a\}, \emptyset \rangle$. We have $U(\emptyset) = \{\top \Rightarrow a\}$ and $U(\{a\}) = \emptyset$. Hence, doing a is better than doing nothing.

The notions U-optimality and minimal U-optimality are properties of decisions that can be used to characterize the type of decision making agents. We define two types of rational agents.

Definition 7. *A BO rational agent is an agent that, confronted with a decision specification DS, selects an U-optimal DS decision. A BO parsimonious agent is a BO rational agent that selects a minimal U-optimal DS decision.*

The logic of belief rules employed in this paper has been called simple-minded throughput (and it has been called out_3^+) in input/output logics [26]. The following example illustrates one of its drawbacks. In Savage's terminology [30], the agent does not obey the sure-thing principle.

Example 7. Let $A = \{a\}$, $W = \{p\}$ and $DS = \langle \emptyset, \emptyset, \{p \Rightarrow a, \neg p \Rightarrow a\}, \emptyset \rangle$. Any decision is an optimal decision. There is no preference for decision $\{a\}$. If p holds then a is obliged, and if p is false then a is obliged. However, the agent cannot infer that a is the optimal decision.

However, in this paper we no longer consider the particular properties of the logic of rules and the logic of decision we have proposed thus far, but we turn to the notion of goals. This concept is introduced in the following section.

3 Goal-Based Normative Decision Theory

In the previous section, we have explained possible decisions of *BO* agents, in the sense of Definition 7, and introduced U-optimality. In this section, we show that every *BO* rational agent can be understood as a goal-based agent [28]. This is done by assuming that the decisions of a *BO* agent are the result of planning of some of its goals. These goals are in turn assumed to be generated by a goal generation mechanism. The question we answer is what are the properties of goals such that, when they are planned based on the belief rules, they result in U-optimal decisions. In particular, we define a characterization of goals such that the decisions that achieve those goals are U-optimal decisions and each U-optimal decision achieves some goal. This result is what we will call a "representation theorem".

3.1 Goal-Based Optimal Decisions

Goal-based decisions in Definition 8 combine decisions in Definition 3 and the notion of goal, which is a set of propositional sentences. Note that a goal set can contain decision variables (which we call to-do goals) as well as parameters (which we call to-be goals).

Definition 8 (Goal-based decision). *Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification and the goal set $G \subseteq L_{AW}$ a set of sentences. A decision d is a G decision iff $E_B(F \cup d) \models_{AW} G$.*

How to define a goal set for a decision specification? We are looking for goal sets *G* which have the desirable property that all *G* decisions are optimal. One way to start is to consider all derivable goals from an initial decision and a *maximal* set of obligations.

Definition 9 (Derivable goal set). *Let $DS = \langle F, B, O, d_0 \rangle$ be a decision specification. A set of formulas $G \subseteq L_{AW}$ is a derivable goal set of DS iff*

$$G = E_{B \cup O'}(F \cup d_0) \setminus Cn_{AW}(E_B(F \cup d_0))$$

where $O' \subseteq O$ is a maximal *(with respect to set inclusion) set such that*

1. *$E_{B \cup O'}(F \cup d_0)$ is consistent and*
2. *there is a DS decision d that is a G decision.*

However, the following proposition shows that for some derivable goal set *G,* not all *G* decisions are U-optimal.

Proposition 7 (U-optimal G decision). *For a derivable goal set G of some decision specification DS, a G decision does not have to be an U-optimal decision.*

Proof. Reconsider the decision specification in Example 4, $DS = \langle \emptyset, \emptyset, \{a \Rightarrow b\}, \emptyset \rangle$. The derivable goal set is $G = \emptyset$. The decision $d = \{a\}$ is a G decision, but it is not U-optimal.

The following proposition shows that the former proposition also holds if we restrict ourselves to minimal optimal decisions.

Proposition 8 (Minimal G decision). *For a derivable goal set G of some decision specification DS, a minimal G decision does not have to be an U-optimal decision.*

Proof. *Consider the decision specification* $DS = \langle \emptyset, \{a \Rightarrow p\}, \{\top \Rightarrow p, a \Rightarrow b\}, \emptyset \rangle$. *The set* $G = \{p\}$ *is the only derivable goal set (based on* $O' = \{\top \Rightarrow p, a \Rightarrow b\}$*). The DS decision* $d_1 = \{a\}$ *is a minimal G decision, but only G decision* $d_2 = \{a, b\}$ *is an U-optimal decision.*

Finally, the following proposition shows that there are also derivable goal sets *G* such that there exist no *G* decision at all.

Proposition 9 (Existence). *For a derivable goal set G of some decision specification DS, G decisions do not have to exist.*

Proof. *Consider the decision specification* $DS = \langle \emptyset, \emptyset, \{\top \Rightarrow p\}, \emptyset \rangle$. *The set* $G = \{p\}$ *is the only derivable goal set (based on* $O' = \{\top \Rightarrow p\}$*). However, the only DS decisions is* $d = \emptyset$ *and G is not a* d *decision.*

Given this variety of problems, we do not try to repair the notion of derivable goal set. Instead, we define goals with respect to an optimal decision.

Definition 10 (Achievable goal set). *Let* $DS = \langle F, B, O, d_0 \rangle$ *be a decision specification. A set of formulas* $G \subseteq L_{AW}$ *is an achievable goal set of DS iff there is an U-optimal DS decision* d *such that*

$$G = \{x \wedge y \mid x \Rightarrow y \in O, E_B(F \cup d) \models_{AW} x \wedge y\}$$

The following two properties show that the notion of achievable goal set does not characterize goals, in the sense that the representation theorem cannot be proven. In particular, the following proposition shows that we can define one half of the representation theorem for achievable goal sets.

Proposition 10. *For an U-optimal decision* d *of DS there is an achievable goal set G of DS such that* d *is a G decision.*

Proof. *Follows directlyfrom the definition.*

However, the following proposition shows that the other half of the representation theorem still fails.

Proposition 11. *For an achievable goal set G ofDS, a G decision does not have to be an U-optimal decision.*

Proof. *Consider specification* $DS = \langle \{\neg q\}, \{a \Rightarrow p, b \Rightarrow p\}, \{\top \Rightarrow p, b \Rightarrow q\}, \emptyset \rangle$. *The set* $G = \{p\}$ *is the only achievable goal set (based on* $O' = \{\top \Rightarrow p, b \Rightarrow q\}$*). The DS decisions* $d_1 = \{a\}$ *and* $d_2 = \{b\}$ *are both (minimal) G decisions, but only* d_1 *is an optimal decision.*

The counter-example in Proposition 11 also shows that we cannot prove the second half of the representation theorem, because we only consider positive goals (states the agent wants to reach) and not negative goals (states the agents wants to evade). The theory is extended with positive and negative goals in the following subsection.

3.2 Positive and Negative Goals

In this section we show that the representation theorem works both ways if we add negative goals, which are defined in the following definition as states the agent has to avoid. They function as constraints on the search process of goal-based decisions.

Definition 11 (Goal-based decision). *Let $DS = \langle F, B, O, d_0\rangle$ be a decision specification, and the so-called positive goal set G^+ and negative goal set G^- subsets of L_{AW}. A decision d is a $\langle G^+, G^-\rangle$ decision iff $E_B(F \cup d) \models_{AW} G^+$ and for each $g \in G^-$ we have $E_B(F \cup d) \not\models_{AW} g$.*

Based on this definition of goal decision, we can extend the definition of achievable goal set with negative goals.

Definition 12 (Positive and negative achievable goal set). *Let $DS = \langle F, B, O, d_0\rangle$ be a decision specification. The two sets of formulas $G^+, G^- \subseteq L_{AW}$ are respectively positive and negative achievable goal sets of DS iff there is an optimal DS decision d such that*

$$G^+ = \{x \wedge y \mid x \Rightarrow y \in O, E_B(F \cup d) \models_{AW} x \wedge y\}$$

$$G^- = \{x \mid x \Rightarrow y \in O, E_B(F \cup d) \not\models_{AW} x\}$$

For $\langle G^+, G^-\rangle$ decisions, we consider minimal optimal decisions. The following example illustrates the distinction between optimal $\langle G^+, G^-\rangle$ decisions and minimal optimal $\langle G^+, G^-\rangle$ decisions.

Example 8. Let $A = \{a, b\}, W = \emptyset$ and $DS = \langle \emptyset, \emptyset, \{a \Rightarrow b\}, \emptyset\rangle$. The optimal decision is $\emptyset$ or $\{a, b\}$, and the related goal sets are $\langle G^+, G^-\rangle = \langle \emptyset, \{a\}\rangle$ and $\langle G^+, G^-\rangle = \langle \{a \wedge b\}, \emptyset\rangle$. The only minimal optimal decision is the former.

The following example illustrates a conflict.

Example 9. Let $W = \{p\}$, $A = \{a\}$, $DS = \langle F, B, O, d_0\rangle$ with $F = \emptyset$, $B = \emptyset$, $O = \{\top \Rightarrow a \wedge p, \top \Rightarrow \neg a\}$, $d_0 = \emptyset$. We have optimal decision $\{\neg a\}$ with goal set $\langle G^+, G^-\rangle = \langle \{\neg a\}, \emptyset\rangle$. The decision $\{a\}$ does not derive goal set $\langle G^+, G^-\rangle = \langle \{a \wedge p\}, \emptyset\rangle$. One of the possible choices is $\{a\}$, which is however sub-optimal since we cannot guarantee that the first obligation is fulfilled.

The following two propositions show that $\langle G^+, G^-\rangle$ goal set is the right characterization of goals such that the representation theorem can be proven. The first part of the representation theorem is analogous to Proposition 10.

Proposition 12. *For an U-optimal decision d of DS there is an achievable goal set $\langle G^+, G^-\rangle$ of DS such that d is a $\langle G^+, G^-\rangle$ decision.*

Proof. *See Proposition 10.*

In contrast to achievable goal set G, the second part of the representation theorem can be proven for an $\langle G^+, G^-\rangle$ goal set.

Proposition 13. *For an achievable goal set $\langle G^+, G^- \rangle$ of DS, a $\langle G^+, G^- \rangle$ decision is an U-optimal decision.*

Proof. *$\langle G^+, G^- \rangle$ is achievable and thus there is an U-optimal DS decision such that $E_B(F \cup d) \models_{AW} G^+$ and for all $g \in G^-$ we have $E_B(F \cup d) \not\models_{AW} g$. Let d be any decision such that $E_B(F \cup d) \models_{AW} G^+$ and for all $g \in G^-$ we have $E_B(F \cup d) \not\models_{AW} g$. Suppose d is not U-optimal. This means that there exists a d' such that $d' >_U d$, i.e., such that $U(d') \subset U(d)$ and there exists an obligation $x \Rightarrow y \in O$ with $E_B(F \cup d) \models_{AW} x$, $E_B(F \cup d) \not\models_{AW} y$ and either $E_B(F \cup d') \not\models_{AW} x \wedge y$ or $E_B(F \cup d') \models_{AW} y$. However, the first option is not possible due to the positive goals and the second option is not possible due to the negative goals. Contradiction, so d has to be U-optimal.*

The representation theorem is a combination of Proposition 12 and 13.

Theorem 1. *A DS decision d is an U-optimal decision if and only if there is an achievable goal set $\langle G^+, G^- \rangle$ of DS such that d is a $\langle G^+, G^- \rangle$ decision.*

The following example illustrates uncertainty about the world.

Example 10. Let $W = \{p, q\}$, $A = \{a\}$, $DS = \langle F, B, O, d_0 \rangle$ with $F = \emptyset$, $B = \{a \Rightarrow q\}$, $O = \{\top \Rightarrow p, p \Rightarrow q\}$, and $d_0 = \emptyset$. We have two optimal decisions, $d_1 = \emptyset$ and $d_2 = \{a\}$, with corresponding achievable goal sets $\langle G^+, G^- \rangle = \langle \emptyset, \{p\} \rangle$ and goal $\langle G^+, G^- \rangle = \langle \{p, q\}, \emptyset \rangle$. We may select $\{a\}$ whereas we do not know whether p will be the case. If we are pessimistic, we assume p will be false. There is no preference to do $\{a\}$.

The following example illustrates side effects from actions.

Example 11. Let $W = \{p, q\}$, $A = \{a\}$, $DS = \langle F, B, O, d_0 \rangle$ with $F = \emptyset$, $B = \{a \Rightarrow p, a \Rightarrow q\}$, $O = \{\top \Rightarrow p, \top \Rightarrow \neg q\}$, and $d_0 = \emptyset$. We have two optimal decisions, $d_1 = \emptyset$ and $d_2 = \{a\}$, with corresponding achievable goal sets $\langle G^+, G^- \rangle = \langle \emptyset, \{p, \neg q\} \rangle$ and goal $\langle G^+, G^- \rangle = \langle \{p, q\}, \emptyset \rangle$. a implies an obligatory proposition, but it also violates another obligation.

The following example illustrates a zig zag. In this example we can continue to construct new goals and new decisions due to side effects of actions.

Example 12. Let $W = \{p, q\}$, $A = \{a\}$, $DS = \langle F, B, O, d_0 \rangle$ with $F = \emptyset$, $B = \{a_i \Rightarrow p_i \mid i = 0, 1, 2, 3, \ldots\}$, $O = \{\top \Rightarrow p_0\} \cup \{a_i \Rightarrow p_{i+1} \mid i = 0, 1, 2, 3, \ldots\}$, and $d_0 = \emptyset$. We have:
$G_0^+ = \{p_0\}$
$d_0 = \{p_0, a_0\}$
$G_1^+ = \{p_0, a_0, p_1\}$
$d_1 = \{p_0, a_0, p_1, a_1\}$
...

4 Agent Specification and Design

In this section we discuss how the proposed qualitative normative decision and goal theory can be used to guide the design and specification of BO rational agents in a compositional way. The general idea of compositional specification and design is to build agents using components. They may be either primitive or composed components, such that the specification of agents can be broken down into the specification of components and their relations. Here we give some preliminary ideas and explain how the proposed qualitative normative decision and goal theory supports a specific compositional design for a BO rational agent.

The qualitative decision theory, as proposed in section 2, specifies the decision making of an agent in terms of its observations and its mental attitudes such as beliefs and obligations. The specified agent can therefore be considered as consisting of components that represent agent's beliefs and obligations and a reasoning component that generates agent's decisions based on its observations and mental attitudes. The abstract design of such a BO agent is illustrated in Figure 1 and copied in Figure 3 below. For this design of BO agents, notions such as optimal decisions and minimal optimal decisions can be used to specify the reasoning component and thus the decision making mechanism of the agent.

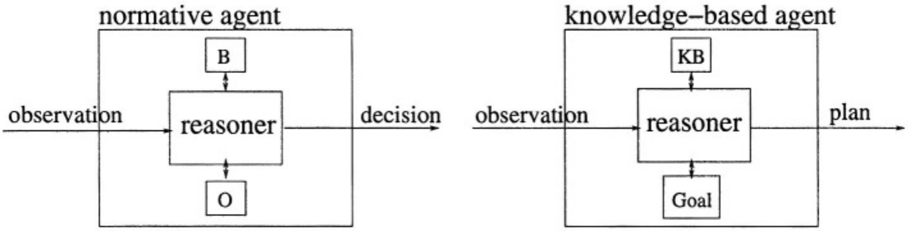


Fig. 3. Agent

The following example illustrates an agent with beliefs and obligations, the possible decisions that the agent can make, and how the notions from qualitative normative decision theory can be used to specify the subset of decisions that the agent can make.

Example 13. Consider an agent who believes that he works and that if he sets an alarm clock he can wake up early to arrive in time at his work,

$$B = \{\top \Rightarrow \mathsf{Work}, \mathsf{SetAlarm} \Rightarrow \mathsf{InTime}\}$$

The agent has also the obligation to arrive early at his work and he has to inform his boss when he does not work:

$$O = \{\mathsf{Work} \Rightarrow \mathsf{InTime}, \neg\mathsf{Work} \Rightarrow \mathsf{InformBoss}\}$$

In this example, the propositions SetAlarm and InformBoss are assumed to be decision variables (the agent has control on setting the alarm clock and informing his boss), while Work and Intime are assumed to be world parameters (the agent has no direct control on its working status and the starting time). Moreover, we assume that the agent has no observation and no intentions. One can specify the agent as a BO rational agent in the sense that it makes optimal decisions. Being specified as a BO rational agent, he will decide to use the alarm clock though he has in principle many possible decisions including $\emptyset$, {SetAlarm}, {InformBoss}, and {SetAlarm, InformBoss}.

The goal-based decision theory, as proposed in section 3, explains the decision making of a BO rational agent as if it aims at maximizing achieved normative goals. In particular, the goal-based decision theory explains how normative goals of an agent can be specified based on its decision specification. The specified reasoning component of the BO rational agent can therefore be decomposed and designed as consisting of two reasoning components: one which generates normative goals and one which generate decisions to achieve those goals. This decomposition suggests an agent design as illustrated in Figure 2 and copied in Figure 4. According to this agent design, a BO agent generates first its normative goals based on its observation, its beliefs, obligations and its intentions. The generated goals are subsequently the input of the decision generation component.

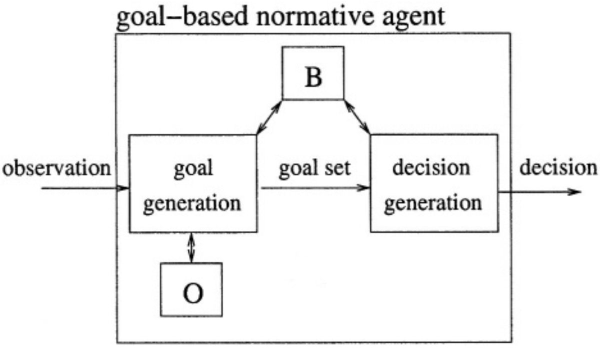


Fig. 4. Goal-based agent

Following the design decomposition, the specification of a BO agent can now also be decomposed and defined in terms of the specification of its goal and decision generation mechanisms. In particular, the goal generation mechanism can be specified in terms of agent's observations and its mental state on the one hand and its goals on the other hand. The decision generation component can then be specified in terms of the agent's goals and mental state on the one hand and its decisions on the other hand.

For example, consider again the working agent that may have in principle many goal sets consisting of $\emptyset$, {Work}, {Intime}, {SetAlarm}, and {InformBoss}. This implies that the goal generation component may generate one of these possible goal sets. Using the notions from goal-based decision theory one may specify the goal generation mechanism in order to generate achievable goal sets which when planned by the decision generation component will result in optimal decisions.

5 Related Research

We draw inspiration from Savage's classical decision theory [30]. The popularity of this theory is due to the demonstration that a rational decision maker, which satisfies some innocent looking properties, acts *as if* it is maximizing its expected utility function. This is called a representation theorem. In other words, Savage does not assume that an agent has a utility function and probability distribution which the agent uses to make decisions. He shows that if an agent bases his decisions on preferences and some properties of these preferences, then we can assume that the agent bases his decisions on these utilities and probabilities together with the decision rule which maximizes its expected utility. Savage therefore does not have to explain what a utility function *is,* an ontological problem which had haunted decision theory for ages.

The theories in Thomason's BDP [33] and Broersen et al.'s BOID [9| are different, because they allow multiple belief sets. This introduces the new problem of blocking wishful thinking discussed extensively in [10]. In earlier work such as [35] we use the set of violated and reached obligations to order states, in the sense that we minimized violations and maximized reached obligations. The present definition has the advantage that it is simpler because it is based on a single minimization process only. Note that in the present circumstances we cannot minimize violations only, because it would lead to the counterintuitive situation that the minimal decision $d = d_0$ is always optimal.

In this paper we have restricted our discussion to beliefs and obligations, and the question can be raised how the decision theory can be extended with desires and intention to the full BOID architecture [9]. Moreover, we have restricted our analysis to a single autonomous agent. However, norms become useful in particular when several agents are considered in a multiagent system. We have made some preliminary observations based on a qualitative game theory in [15,17].

Boella and Lesmo [2] introduce sanction-based norms in a model in which the normative system itself is modeled as an agent, and decision-making in the context of norms becomes playing a game with the normative agent. Their model can be motivated by an attribution of mental attitudes to autonomous normative systems, which itself can be motivated by social delegation of shared goals to the system, see [3,4,5]. In this paper we have not specified any details of the obligation rules, but an extension of the theory in this paper along the Boella-Lesmo proposal can be found in [6,7].

Conte and Dignum [12] argue that, if you are speaking of normative agents as systems that somehow 'process' norms and decide upon them, then they must first form beliefs about those norms, whether they then adopt the norms or not. We believe that this is not incompatible with the approach advocated in this paper. However, in our general theory we do not want to commit ourselves to this particular view on norms. Our theory can also be applied, for example, to the game-theoretic notion of norms as advocated by for example[31].

A distinction has been made between goal generating norms and action filtering norms (Castelfranchi and Conte, personal communication). It is an open problem how these two kinds of norms can be formalized by our decision theory. It seems that obligation rules are only used to generate goals, and we therefore need another type of norms which filters actions. However, obligations with a decision variable in the head seem to act also as a kind of action filters.

6 Concluding Remarks

In this paper we have given an interpretation for goals in a qualitative decision theory based on beliefs and obligation rules, and we have shown that any agent which makes optimal decisions acts as if it is maximizing its achieved goals. Inspired by Savage, we develop a qualitative normative decision theory in which a normative agent acts *as if* it is trying to maximize achieved normative goals. This is what we call a goal-based representation theorem. It implies that agents can be formalized or verified as goal-based reasoners even when the agent does not reason with goals at all. In other words, goal-based representations do not have to be descriptive. A consequence of this indirect definition of goals is that the theory tells us what a goal *is,* such that we do not have to explain its ontological status separately. We call an agent which minimizes its unreached obligations a BO rational agent, and we define goals as a set of formulas which can be derived by beliefs and obligations in a certain way. Our central result thus says that *BO rational agents act as if they maximize the set of goals that will be achieved.*

We believe that the qualitative normative decision theory and goal-based decision theory can be used to provide compositional specification and design of BO rational agents. This leads to a transparent agent specification and design structure. Moreover, it leads to support for reuse and maintainability of components and generic models. The compositional specification and design of agents enable us to specify and design agents at various levels of abstraction leaving out many details such as representation issues and reasoning schemes. For our BO rational agents we did not to explain how decisions are generated; we only specified what decisions should be generated. At one lower level we decomposed the reasoning mechanism and specified goal and decision generation mechanisms. We also did not discuss the representation of individual components such as the belief or the obligation components. The conditional rules in these components specify the input/output relation.

Our motivation comes from the analysis of goal-based architectures, which have recently been introduced. However, the results of this paper may be relevant for a much wider audience. For example, Dennett argues that automated systems can be analyzed using concepts from folk psychology like beliefs, obligations, and goals. Our work may be used in the formal foundations of this ‘intentional stance’ [18].

There are several topics for further research. The most interesting question is whether belief and obligation rules are fundamental, or whether they in turn can be represented by some other construct. Other topics for further research are a generalization of our representation theorem to other choices in our theory, the development of an incremental approach to goals, and the development of computationally attractive fragments of the logic, and heuristics of the optimization problem.

Acknowledgements. Thanks to Guido Boella, Zhisheng Huang and Joris Hulstijn for discussions on the issues raised in this paper.

References

1. J.R. Anderson, M. Matessa, and C. Lebiere. ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction,* 12(4):439–462, 1997.
2. G. Boella and L. Lesmo. A game theoretic approach to norms. *Cognitive Science Quarterly,* 2(3-4):492–512, 2002.
3. G. Boella and L. van der Torre. Attributing mental attitudes to normative systems. In *Proceedings of the second international joint conference on autonomous agents and multi agent systems (AAMAS'03),* 2003.
4. G. Boella and L. van der Torre. Rational norm creation: Attributing mental states to normative systems, part 2. In *Proceedings of the Eighth International Conference on Artificial Intelligence and Law (ICAIL'03).* ACM, 2003.
5. G. Boella and L. van der Torre. Decentralized control: Obligations and permissions in virtual communities of agents. In *Proceedings of the Fourteenth International Symposium on Methodologies for Intelligent Systems (ISMIS'03).* Springer, 2003.
6. G. Boella and L. van der Torre. Norm governed multiagent systems: The delegation of control to autonomous agents. In *Proceedings of the 2003 IEEE/WIC International Conference on Intelligent Agent Technology (IAT'03).* IEEE, 2003.
7. G. Boella and L. van der Torre. Local policies for the control of virtual communities. In *Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence (WI'03).* IEEE, 2003.
8. C. Boutilier. Toward a logic for qualitative decision theory. In J. Doyle, E. Sandewall, and P. Torasso, editors, *Principles of Knowledge Representation and Reasoning, Proceedings of the Fourth International Conference (KR'94),* pages 75–86, San Francisco, California, 1994. Morgan Kaufmann Publishers.
9. J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly,* 2(3-4):428–447, 2002.
10. J. Broersen, M. Dastani, and L. van der Torre. Realistic desires. *Journal of Applied Non-Classical Logics,* 12(2):287–308, 2002.
11. P.R. Cohen and H.J. Levesque. Intention is choice with commitment. *Artificial Intelligence,* 42:213–261, 1990.
12. R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In J. Muller, M. Singh, and A. Rao, editors, *Intelligent Agents V (ATAL'98),* volume 1555 of *LNAI,* pages 319–333. Springer, 1999.
13. M. Dastani, Z. Huang, and L. van der Torre. Dynamic desires. In S. Parsons, P. Gmytrasiewicz, and M. Wooldridge, editors, *Game Theory and Decision Theory in Agent-Based Computing,* volume 5 of *Multiagent Systems, Artificial Societies and Simulated Organizations,* pages 65–79. Kluwer, 2002.
14. M. Dastani, J. Hulstijn, and L. van der Torre. How to decide what to do? *European Journal of Operations Research,* to appear.
15. M. Dastani and L. van der Torre. Decisions and games for BD agents. In *Proceedings GTDT'02, Papers from the AAAI workshop,* pages 37–43. AAAI Press, 2002. Technical report WS-02-06.
16. M. Dastani and L. van der Torre. Specifying the merging of desires into goals in the context of beliefs. In *Proceedings of The First Eurasian Conference on Advances in Information and Communication Technology (EurAsia ICT 2002),* LNCS. Springer, 2002.

17. M. Dastani and L. van der Torre. What is a joint goal? Games with beliefs and defeasible desires. In *Proceedings of Ninth International Workshop on Non-Monotonic Reasoning (NMR'02),* 2002.
18. D. Dennett. *The intentional stance.* MIT Press, Cambridge, MA, 1987.
19. F. Dignum, D. Morley, E. Sonenberg, and L. Cavedon. Towards socially sophisticated bdi agents. In *Proceedings of the fourth International Conference on MultiAgent Systems (ICMAS-2000),* pages 111–118, Boston, 2000. IEEE Computer Society.
20. J. Doyle. A model for deliberation, action and introspection. Technical Report AI-TR-581, MIT AI Laboratory, 1980.
21. M. Georgeff and A. Lansky. Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence,* pages 677–682, 1987.
22. S. Kraus, K. Sycara, and A. Evenchik. Reaching agreements through argumentation: a logical model and implementation. *Articial Intelligence,* 104:1–69,1998.
23. J.E. Laird, A. Newell, and P.S. Rosenbloom. SOAR: An architecture for general intelligence. *Artificial Intelligence,* 33(1):1–64, 1987.
24. J. Lang. Conditional desires and utilities - an alternative approach to qualitative decision theory. In *In Proceedings of the European Conference on Artificial Intelligence (ECAI'96),* pages 318–322, 1996.
25. J. Lang, L. van der Torre, and E. Weydert. Utilitarian desires. *Autonomous Agents and Multi-Agent Systems,* 5:3:329–363, 2002.
26. D. Makinson and L. van der Torre. Input-output logics. *Journal of Philosophical Logic,* 29:383–408, 2000.
27. D. Makinson and L. van der Torre. Constraints for input-output logics. *Journal of Philosophical Logic,* 30:155–185, 2001.
28. A. Newell. The knowledge level. *Artificial Intelligence,* 18(1):87–127, 1982.
29. A. S. Rao and M. P. Georgeff. Decision procedures for BDI logics. *Journal of Logic and Computation,* 8:293–342, 1998.
30. L. Savage. *Thefoundations ofstatistics.* Wiley, New York, 1954.
31. Y. Shoham and M. Tennenholtz. On the emergence of social conventions: Modeling, analysis, and simulations. *Artificial Intelligence,* 94:139–166, 1997.
32. H. A. Simon. *The Sciences of the Artificial.* MIT Press, Cambridge, MA, second edition, 1981.
33. R. Thomason. Desires and defaults: A framework for planning with inferred goals. In *Proceedings of Seventh International Conference on Knowledge Representation and Reasoning (KR'00),* pages 702–713, 2000.
34. L. van der Torre. Contextual deontic logic: Normative agents, violations and independence. *Annals of Mathematics and Artificial Intelligence,* 37 (1-2):33–63, 2003.
35. L. van der Torre and Y. Tan. Diagnosis and decision making in normative reasoning. *Artificial Intelligence and Law,* 7(1):51–67, 1999.
36. L. van der Torre and E. Weydert. Parameters for utilitarian desires in a qualitative decision theory. *Applied Intelligence,* 14:285–301, 2001.
37. G.H. von Wright. *Practical Reason: Philosophical papers, volume 1.* Basil Blackwell, Oxford, 1983.

Searching for a Soulmate – Searching for Tag-Similar Partners Evolves and Supports Specialization in Groups

David Hales

The Centre for Policy Modelling, The Business School, Manchester Metropolitan University,
Manchester, UK.
dave@davidhales.com

Abstract. In a previous paper [1] we presented simulation results that demonstrated the evolution of "tag based" groups composed of cooperative (in-group altruistic) individual agents performing *specialised* functions. We showed how "teams" of individual maximisers (who copy the behaviours of those who outperform them) come to form internally specialised and cooperative groups that efficiently exploit their environment. We have also demonstrated [1, 2, 3] that the efficiency of the specialisation process is highly dependent on the "searching strategy" employed by agents to locate in-group members with required skills. Specifically we showed that populations of agents with "smart" searching strategies outperformed populations of "dumb" (random) search strategies – even when the costs of smart searching were much higher. We hypothesised that in mixed populations smart strategies would out-evolve dumb ones. In this paper we test this hypothesis. Our results show that smart strategies do indeed outperform dumb strategies for significant periods of time but that dumb strategies persist also. The time series of individual runs show cycles of smart and dumb strategies in the population over generations. We argue that the study of such phenomena offers a possible minimal way towards understanding the evolution of institutional roles and internal specialisation – without positing actions that originate at the supra-individual level (though we do not discount such actions).

1 Introduction

A previous model [1] demonstrated tag[1] processes that were sufficient to evolve sustained altruistic behaviour between specialised agents. Those results present a new way to address an old puzzle [5] – why would self-interested *(Homo economicus)* or "selfish-gene" evolutionary adaptive agents help each other altruistically (i.e. perform some behaviour that reduces their utility or fitness but increases that of another)? The tag-based mechanism presented was not dependant on any kind of direct

[1] Tags are identifiable markers (physical markings, gestures or social cues) associated with particular agents that other agents can observe. See Holland [4] for some early speculation on the potential power of tags for capturing processes of emergence via social interactions.

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 228–239, 2004.

reciprocation based on repeated interactions [6]. An earlier model [7] applied the same mechanism to the one-shot dyadic Prisoners Dilemma (PD) game demonstrating that very high levels of cooperation were possible. By extending this model we investigated if a similar tag process could support the formation of groups of agents with internal specialisation. We found that it could but only partially [1].

The evolution of group-functional behaviours (involving altruism and specialisation) is of interest to the human and biological sciences [8, 9, 10] and, interestingly, to the Multi-Agent System (MAS) engineering community [11, 12, 13, 14]. MAS engineers want to build systems of computational agents that work together to solve problems whereas social and biological scientists want to generate new theories to help understand the natural and social worlds. With these two areas of inquiry in mind we provide results from computational simulation experiments in the artificial domain [15, 16].

We demonstrated [2] that populations composed of agents with *smart* partner selection strategies (where agents search the population for partners with the same tags) outperformed populations of *dumb* selection strategies (where partners are chosen at random from the population). We hypothesised that in a *mixed* population composed of smart *and* dumb strategies the smart strategies would out-evolve the dumb ones.

In this paper we test the hypothesis by implementing search strategies as an evolvable binary trait (representing smart or dumb). We show that even when the costs of the smart strategies are significantly higher than dumb strategies, smart strategies often out-perform dumb strategies and persist in the population. However, they do not eliminate dumb strategies completely. What we find is that the proportions of each strategy in the population change in cycles over time. We offer an explanation of this process with reference to the model dynamics. Essentially, the success of smart strategies ultimately leads to a population in which they are not required. This leads to invasion by dumb strategies which in-turn creates the conditions under which smart strategies will out-evolve dumb strategies again. The result is an oscillating dumb / smart strategy mix. Even though smart strategies *do not* eliminate dumb strategies, the selective pressure for smart strategies is enough to sustain them resulting in a *significant increase* in altruistic and cooperative behaviour (producing more efficient, specialised agent groups) than is the case when only dumb strategies are implemented.

2 The Model

The model consists of a population of 100 evolving agents. The tag matching mechanisms follow that of Riolo et al [17]. The specialisation process follows Hales [1] and the smart and dumb searching strategies follow Hales [2] but are here represented as an evolvable binary trait rather than being hard-coded and fixed

We now briefly summarise the model. Each agent has four traits: a tag ≥ ≥ [0..1], a tolerance threshold $1 \leq T \geq 0$, a skill type $S \geq \{1, 2\}$ and a search strategy $Z \geq$ {smart, dumb}. Initially, tags, thresholds, skills and strategies are allocated uniformly randomly. In each generation, each agent is awarded some number P of resources. Each resource is assigned a required skill type. Resources can only be "harvested" by agents possessing the required skill type. A resource can be interpreted with a biological analogy as some kind of food energy (requiring a certain skill or ability to prepare) or as a computational job (requiring some computational skill or resource to process). The skill type assigned to a resource is randomly selected from those skills that do not match the receiving agents skill[2]. An agent therefore is never awarded a resource that matches its skill type. Since the agents cannot directly harvest resources they search the population for another agent with required skill and tag values.

Donation of a resource occurs if a recipient is found with the required skill type and with a *sufficiently similar* tag value. A recipient tag is considered to be sufficiently similar if it is within the tolerance of the donating agent. Specifically, given a potential donor agent D and a potential recipient R a donation will only be made when $|\geq_D - \geq_R| \geq T_D$. This means that an agent with a high T value may donate to agents over a large range of tag values. A low value for T restricts donation to agents with very similar tag values to the donor. In all cases donation can only occur when the skill type of the receiving agent matches the skill type associated with the resource. If a donation is made the donating agent incurs a cost, c, and the recipient gains a benefit, b (since it can harvest the resource). In all experiments given in this paper, the benefit b = 1. The cost, c, depends on the value of the search strategy Z. When Z = smart, c = 0.5, when Z = dumb then c = 0.1. Here we capture the notion that "smart searching" is more costly than dumb searching. Figure 1 shows schematically how resources might be passed.

If an agent uses a dumb search strategy the search simply involves a single random selection from the population. If the randomly selected partner does not have the correct tag and skill type then no donation takes place. In contrast, if an agent uses a smart strategy it searches the entire population for a potential recipient with appropriate tag and skill values. Here, we do not model the *actual mechanism employed* but just the *outcome* assuming a smart strategy were used[3]. We assume that some efficient mechanism exists which allows agents to *find a potential recipient in the population if one exists*[4]. As discussed previously [1] a number of plausible mechanisms can be hypothesised – based on spatial and/or cognitive relationships (e.g. "small world" social networks [18], meeting places, central stores [19] etc.).

After all agents have been awarded P resources and made any possible donations the entire population is reproduced. Reproduction is accomplished in the following manner – each agent is selected from the population in turn, its score is compared to another randomly chosen agent, and the agent with the highest score is reproduced.

[2] Results obtained from a model in which agents may be awarded resources matching their own skill types produced similar results to those presented in this paper.

[3] Moreover we only allow our evolutionary process to select a binary dumb / smart trait. In any physically instantiated society (human, animal, robotic) it would seem that many possible strategies would be available. Our aim here is to show minimally that non-random mixing strategies *can* evolve.

[4] If several suitable recipients exist in the population we assume here that one of them is selected to receive the donation at random.

Mutation is applied to each trait of each offspring. With probability 0.1 the offspring receives a new tag (uniformly randomly selected). With the same probability, Gaussian noise is added to the tolerance value (mean 0, standard deviation 0.01). When T < 0 or T > 1, it is reset to 0 and 1 respectively. Also with probability 0.1 the offspring is given a new skill type (uniformly randomly selected) and with the same probability the search strategy is
changed.

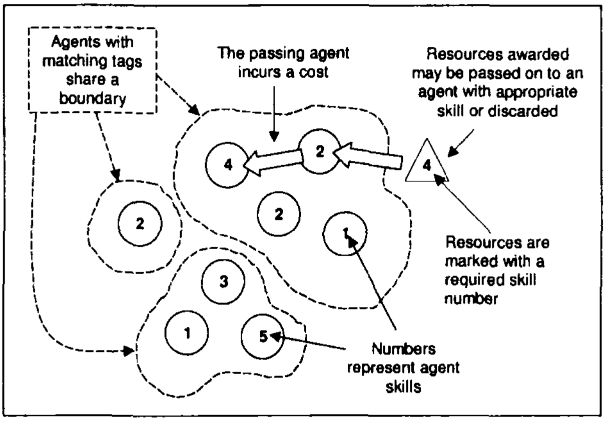

Agents with matching tags share a boundary
The passing agent incurs a cost
Resources awarded may be passed on to an agent with appropriate skill or discarded
4
2
4
2
2
1
3
Resources are marked with a required skill number
1
5
Numbers represent agent skills

Fig. 1. A schematic representation of how a resource may be passed to an in-group with the required skill (at a cost to the passing agent).

3 Results

The first set of results, in Table 1, below, show the donation rates achieved as a percentage of total awards made, the average tolerance values and the average number of smart agents in a 2-skill scenario. The shaded columns are results from previous papers [1,2] and are given for comparison purposes.

All results are over 30,000 generations with 30 replications. Each replication represents an individual run started with a different pseudo-random number seed. The standard deviations are over the 30 runs executed for each unique P value setting[5] and strategy. The column labelled "dumb" shows results from previous experiments using a *dumb* random recipient search strategy [1]. In this condition all agents in the population use a dumb random search strategy. The columns labelled "smart" show results when all agents use a *smart* strategy [2].

[5] The standard deviations are not calculated over the percentages given but proportions (i.e. percentages scaled within [0..1] – so 100% would count as 1 and 50% as 0.5 etc.)

Table 1. Donation rates and tolerance levels for different numbers of awards in a 2-skill scenario (i.e. when S ε {1,2}) and for different search strategies. The values in brackets are standard deviations over the 30 replications. The shaded columns are results from pervious papers [1, 2] given for comparison purposes. The P column gives the number of resource awards. The Z column gives the average proportion of smart (searching) agents in the population (averaged over each generation).

	Donation Rate – Ave %			Tolerance – Ave			
P	Dumb	Mixed·	Smart	Dumb	Mixed	Smart	Z
1	1.5	11.7	29.5	0.028	0.007	0.021	0.358
	(0.001)	(0.001)	(0.081)	(0.002)	(0.001)	(0.084)	(0.002)
2	1.1	13.9	47.9	0.019	0.006	0.030	0.320
	(0.000)	(0.002)	(0.087)	(0.001)	(0.000)	(0.111)	(0.001)
3	1.0	25.3	59.9	0.015	0.010	0.017	0.230
	(0.000)	(0.004)	(0.035)	(0.001)	(0.000)	(0.048)	(0.002)
4	0.9	37.7	66.3	0.013	0.014	0.028	0.193
	(0.000)	(0.002)	(0.046)	(0.000)	(0.005)	(0.105)	(0.001)
6	0.9	41.9	70.9	0.011	0.017	0.011	0.165
	(0.000)	(0.001)	(0.010)	(0.001)	(0.004)	(0.014)	(0.000)
8	0.9	42.9	73.1	0.010	0.018	0.009	0.151
	(0.000)	(0.001)	(0.002)	(0.000)	(0.001)	(0.000)	(0.000)
10	2.1	43.4	74.5	0.010	0.020	0.009	0.143
	(0.002)	(0.001)	(0.002)	(0.000)	(0.001)	(0.000)	(0.000)
20	12.9	43.6	77.3	0.025	0.023	0.010	0.130
	(0.000)	(0.001)	(0.002)	(0.003)	(0.004)	(0.001)	(0.000)
40	13.9	43.5	79.3	0.098	0.023	0.038	0.125
	(0.015)	(0.001)	(0.033)	(0.190)	(0.004)	(0.107)	(0.000)

The "mixed" columns give the new results obtained when the searching strategy is allowed to evolve as a binary trait (as described above). The Z column gives the average proportion of smart agents in the population (averaged over each generation). So if this value were 1 it would mean that at the end of each generation (for 30,000 generations) all agents were using smart searching strategies (i.e. Z=smart for all agents at all times). The values in this column are only meaningful with reference to the associated "mixed" columns. Additionally, these values give no indication of the time dynamics of strategy evolution. In all cases the cost to a donor agent of using a dumb strategy is c = 0.1, for smart it is c = 0.5. The benefit, b, passed to any recipient of a donation is held at 1.

Figures 2 and 3 present some of the data from table 1 (the 2-skill scenario) in graphical form. Figure 2 shows the donation rates for dumb, smart and mixed populations. Note that for mixed populations the donation rate is *significantly higher* than for dumb populations. This indicates that smart strategies are being selected-for by the evolutionary process and that this improves donation rates. However, the donation rate is *significantly lower* than when the population is set to all smart – indicating that smart strategies are not completely dominating the population and stabilising. Figure 3 shows the donation proportion for the mixed case along with the average proportion of the population using the smart search strategy. This illustrates clearly that as the number of rewards given to each agent increases the proportion of agents holding smart strategies decreases.

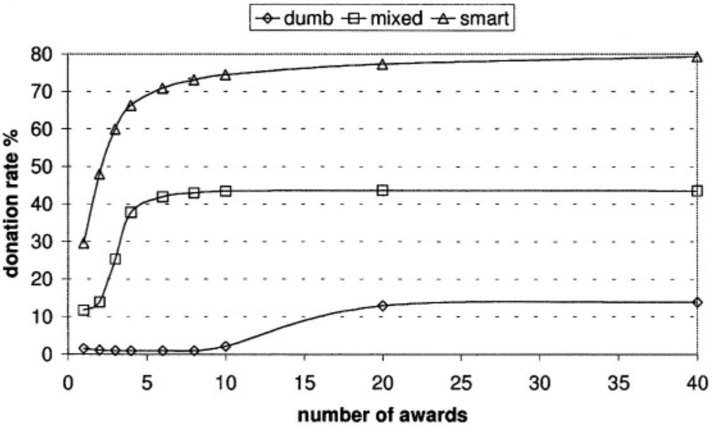


Fig. 2. Comparison of donation rates for dumb, smart and mixed populations for different numbers of resource rewards to each agent when skills = 2. Notice the significant increase in donation rates in the mixed population relative to the dumb population. This indicates there is some selective pressure for smart strategies.

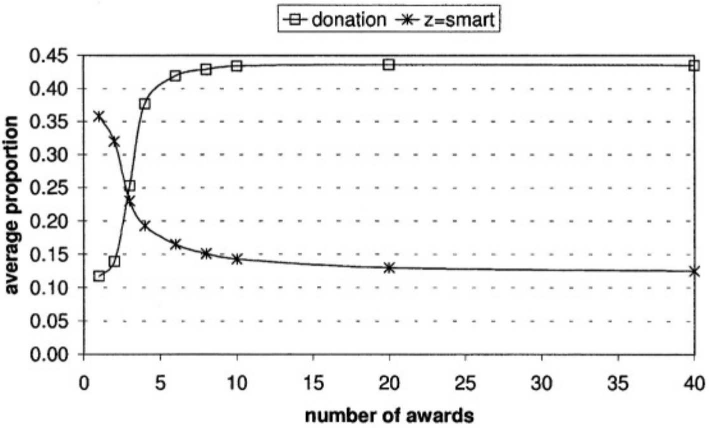


Fig. 3. Average proportion of smart agents in the population over all generations for different numbers of resource awards for 2-skill scenario. For comparison purposes the donation rate for the mixed case is here included (mapped as a proportion rather than percentage as shown in figure 2). Notice that as the number of awards increases the proportion of smart agents decreases but does not go to zero.

This can be explained as follows: when many awards are being made, agents using the dumb strategy have a higher chance of finding at least some matching partners for donation. Conversely, the smart strategy agents will always find a matching partner if one exists – however the costs to the smart agent are higher. When an in-group of agents reaches a significant size (generally sharing identical tags) then there will be little advantage agents in that group when using a smart strategy – since dumb searching will be almost as good but costs much less.

However, why don't smart strategies dominate the population when the number of rewards is low? To answer this question we advance an explanation that produces a hypothesis concerning the dynamics of strategy evolution. When a large number of smart strategies predominate in the population, agents, practicing dumb strategies, have an opportunity to "free-ride". The smart agents will incur a large cost locating and donating to dumb agents but this will not be reciprocated. Given this, why don't smart strategies disappear completely? They do not disappear because tag mutation and low tolerance (notice in table 1 tolerance is low when rewards are low) keep a diversity of "in-groups" (generally all sharing the same tag value) in the population – i.e. a diversity of tags. Given this, "tag groups" which become invaded and dominated by dumb strategies, become dysfunctional. The dumb invaders sow the seeds of their own destruction. By exploiting the group, they kill it. In this situation we expect other tag groups that are composed entirely of smart agents to outperform the dysfunction groups. This is the same process that allows the specialisation to evolve [1]. The constant formation of new tag groups produces selective pressure for group level adaptations. The way that such tag based group processes can suppress free-

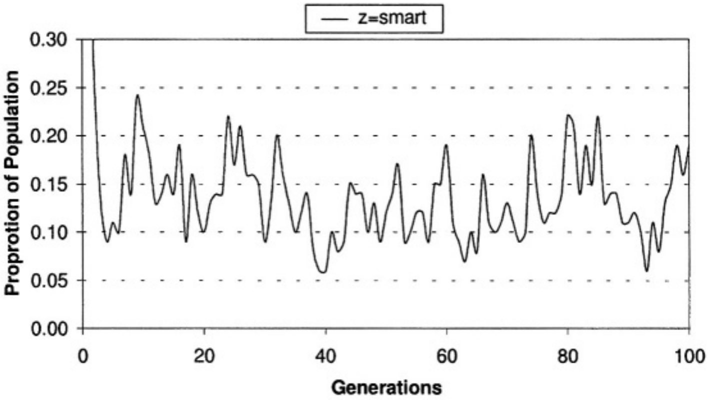


Fig. 4. Proportion of smart agents over 100 generations for a typical run. Number of awards = 10, number skills = 2. The proportion oscillates over time. The oscillation appears to be occurring on several scales. The initially high value (not fully shown on the scale) is due to the uniformly random initialisation (from which we expect 50% of agents to be initialised as smart).

riding within a group is illustrated starkly by Hales [7] where a similar model is applied to the one-shot prisoner's dilemma[6].

From the above analysis we hypothesize that the time dynamics (over generations) of strategy evolution should show a cyclical signature (i.e. phases of high and low proportions of smart strategies over time). Figure 4 shows the proportion of smart agents in the population over 100 generations of a typical run for the 2-skill scenario when awards (P) = 10. As can be seen, the proportion of smart agents does not stay constant over time but oscillates on several scales.

Results from the 5-skill scenario are given in table 2 and figures 5 and 6. The results are qualitatively similar to the 2-skill scenario but the increase in donation rates over the "all dumb" population is less spectacular (though significant).

Table 2. *Donation rates and tolerance levels for different numbers of awards in a 5-skill* **scenario (i.e. there are 5 skill types, such that each agent has a skill S ε {1,2,3,4,5}) and for different search strategies. The values in brackets are standard deviations over the 30 replications. The shaded columns are results from pervious papers [1, 2] given for comparison purposes. The P column shows the number of resource awards. The Z column gives the average proportion of smart (searching) agents in the population (averaged over each generation).**

	Donation Rate – Ave %			Tolerance – Ave			
P	Dumb	Mixed	Smart	Dumb	Mixed	Smart	Z
1	1.5	6.5	29.5	0.028	0.008	0.021	0.395
	(0.001)	(0.001)	(0.081)	(0.002)	(0.001)	(0.084)	(0.002)
2	1.1	5.9	47.9	0.019	0.007	0.030	0.357
	(0.000)	(0.001)	(0.087)	(0.001)	(0.001)	(0.111)	(0.002)
3	1.0	6.1	59.9	0.015	0.007	0.017	0.325
	(0.000)	(0.001)	(0.035)	(0.001)	(0.001)	(0.048)	(0.002)
4	0.9	7.4	66.3	0.013	0.007	0.028	0.283
	(0.000)	(0.002)	(0.046)	(0.000)	(0.000)	(0.105)	(0.002)
6	0.9	15.6	70.9	0.011	0.012	0.011	0.170
	(0.000)	(0.002)	(0.010)	(0.001)	(0.002)	(0.014)	(0.002)
8	0.9	18.8	73.1	0.010	0.014	0.009	0.141
	(0.000)	(0.001)	(0.002)	(0.000)	(0.000)	(0.000)	(0.001)
10	2.1	20.0	74.5	0.010	0.015	0.009	0.130
	(0.002)	(0.001)	(0.002)	(0.000)	(0.001)	(0.000)	(0.000)
20	12.9	21.6	77.3	0.025	0.023	0.010	0.117
	(0.000)	(0.006)	(0.002)	(0.003)	(0.013)	(0.001)	(0.000)
40	13.9	22.0	79.3	0.098	0.034	0.038	0.114
	(0.015)	(0.006)	(0.033)	(0.190)	(0.057)	(0.107)	(0.000)

[6] For a more detailed treatment of the one-shot prisoner's dilemma model and some experimentation with a more sophisticated cultural evolutionary model along with a detailed treatment of this kind of group selective process see Hales [20].

Interestingly, the proportion of smart agents in the population over different reward values was found to be very similar to the 2-skill scenario (compare figures 3 and 6). So, the more modest increase in cooperation obtained in the 5-skill scenario is not a result of less agents evolving smart strategies. This indicates that when groups are required to be more specialised this degrades the effect of smart strategies when they do not dominate the population.

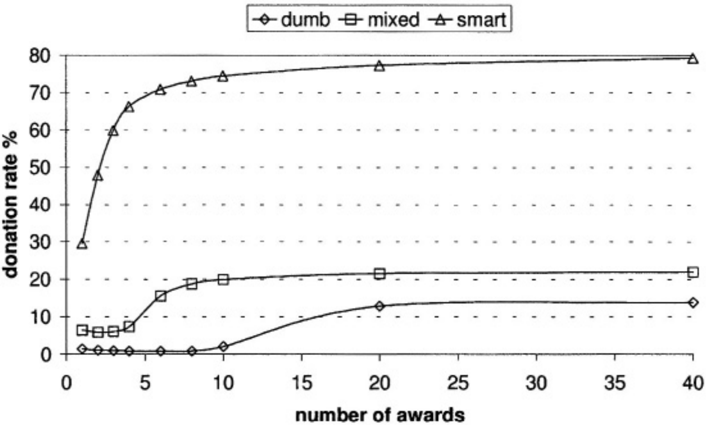


Fig. 5. Donation rates for dumb, smart and mixed populations for different numbers of awards in the 5-skill scenario. The values used for the chart are taken from table 2.

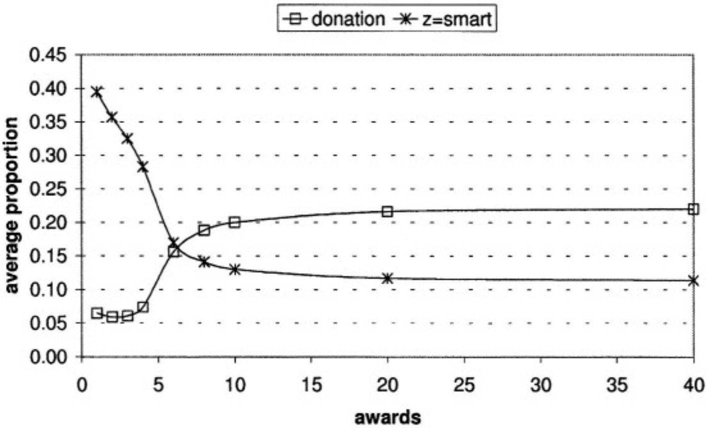


Fig. 6. Average proportion of smart agents in the population over all generations for different numbers of resource awards for 5-skills. For comparison purposes the donation rate for the mixed case is here included (mapped as a proportion rather than percentage as shown in figure 2). Notice that as the number of awards increases the proportion of smart agents decreases but does not go to zero. A very similar signature is given for 2-skills (see figure 3).

4 Discussion

The major conclusion of the paper is that *smart partner searching strategies (even when significantly more costly than dumb strategies) derive enough evolutionary advantage to persist in a population of agents. They result in a significant improvement in the amount of altruistic behaviour produced and hence help to support in-group specialisation.*

However, in the model presented here, smart strategies do not completely dominate the population or even takeover a majority of the population. Additionally, high degrees of in-group specialisation do not appear to benefit as much from similar proportions of smart agents in the population.

We advance the model as an *existence proof* in the artificial domain that *non-random interactions between agents can be selected for by a simple evolutionary process*[7]. Given this, models should not discount non-random mixing – especially in the context of models of complex social behaviours.

Biologically inspired models [17] and culturally oriented models (based on the replicator dynamics) often assume random mixing of populations of agents [21]. However, it would seem that when agents in such models are meant to represent individuals with even rudimentary abilities to learn and distinguish between others based on some evolvable observable characteristic then such an assumption is hard to support. In the model presented in this paper we evolve behaviours far away from simple random mixing. When applied to the understanding of human-like societies assumptions of random mixing preclude the analysis of the evolution of complex social structures.

The large selective advantage to non-random mixing (locating one's in-group) indicates a possible trajectory for the evolution of relatively sophisticated socio-cognitive mechanisms. Starting from locating and distinguishing in-group from out-group it would seem that more sophisticated social behaviours, such as norm enforcement via ostracism (or direct punishment) and the sharing and distribution of collective goods and costs, could also become evolutionarily tenable [10, 19]

We have already demonstrated the evolution of primitive in-group "roles" linked to resource specialisation. However, with more sophisticated group processes we might expect more refined roles to be possible – such as group organisational and informational roles (not directly linked to resource harvesting). In order to model such processes we need to move away from simple "dyadic" two-agent interactions of donation to something more like group level sharing.

We suggest the kinds of processes outlined in this paper may explain the emergence of *proto-institutions.* It would seem that these minimal forms of social organisation would be a necessary condition for the formation of richer and more complexinstitutionalorganisations.

[7] This "existence proof" would be significantly strengthened by further experiments varying a number of parameters – specifically the mutation rate.

Acknowledgements. Work presented here has greatly benefited from discussions with, and ideas from, Bruce Edmunds, Centre for Policy Modelling, Manchester Metropolitan University, UK. Thanks are also due to the anonymous referees from whom encouragement, corrections and suggestions were gratefully received.

References

1. Hales, D. (accepted) Evolving Specialisation, Altruism and Group-Level Optimisation Using Tags. In *Sichman, J. S., Bousquet, F. Davidsson, P. (eds.) Multi-Agent-Based Simulation II. Lecture Notes in Artificial Intelligence 2581.* Berlin: Springer-Verlag (available at www.davidhales.com). (2002).
2. Hales, D.: Smart Agents Don't Need Kin – Evolving Specialisation and Cooperation with Tags. *CPM Working Paper 02-89.* The Centre for Policy Modelling, Manchester Metropolitan University, Manchester, UK (available at: http://www.cpm.mmu.ac.uk/cpmreps.html), (2002).
3. Hales, D.: Wise-Up! - Smart Tag Pairing Evolves and Persists. *CPM Working Paper 02-90.* The Centre for Policy Modelling, Manchester Metropolitan University, Manchester, UK (available at: http://www.cpm.mmu.ac.uk/cpmreps.html), (2002).
4. Holland, J.: The Effect of Labels (Tags) on Social Interactions. *SFI Working Paper 93-10-064.* Santa Fe Institute, Santa Fe, NM, (1993)
5. Hardin, G. The tragedy of the commons. Science, 162:1243–1248, (1968).
6. Trivers, R. L.: The evolution of reciprocal altruism. *Q. Rev. Biol.* 46:35–57 (1971)
7. Hales, D.: Cooperation without Space or Memory: Tags, Groups and the Prisoner's Dilemma. In Moss, S., Davidsson, P. (Eds.) *Multi-Agent-Based Simulation.* Lecture Notes in Artificial Intelligence 1979. Berlin: Springer-Verlag (2000)
8. Soltis, J., Boyd, R. and Richerson, P.: Can Group-functional Behaviours Evolve by Cultural Group Selection? *Current Anthropology,* 36(3). (1995)
9. Wilson, D. S. and Sober, E.: Reintroducing group selection to the human and behavioural sciences. *Behavioural and Brain Sciences,* 17(4):585–654. (1994).
10. Bowles, S. and Gintis, H.: Optimal Parochialism: The Dynamics of Trust and Exclusion in Networks. *SFI Working Paper 00-03-017.* Santa Fe Institute, Santa Fe, N.M. (2000).
11. Hogg, L. M., and Jennings, N. R.: Socially Rational Agents. Proc. *AAAI Fall symposium on Socially Intelligent Agents,* Boston, Mass., November 8-10, 61–63, 1997.
12. Jennings, N., and Campos, J.: Towards a Social Level Characterization of Socially Responsible Agents. *IEE Proceedings on Software Engineering.* 144(1): 11–25, 1997.
13. Kalenka, S.: *Modelling Social Interaction Attitudes in Multi-Agent Systems.* Ph.D Thesis. Department of Electronics and Computer Science, Southampton University, (2001).
14. Kalenka, S., and Jennings, N.R.: Socially Responsible Decision Making by Autonomous Agents. *Cognition, Agency and Rationality* (eds. Korta, K., Sosa, E., Arrazola, X.) Kluwer 135–149, (1999).
15. Gilbert, N. and Doran J., (eds.): *Simulating Societies: the Computer Simulation of Social Phenomena.* London: UCL Press, (1994).
16. Gilbert, N. and Conte, R. (eds.): *Artificial Societies: the Computer Simulation of Social Life, London:* UCL Press, (1995).
17. Riolo, R., Cohen, M. & Axelrod, R.: Cooperation without Reciprocity. *Nature* 414, 441–443. (2001).

18. Watts, J. D.: *Small Worlds – The Dynamics of Networks between Order and Randomness.* Princeton, New Jersey, USA. (1999).
19. Pedone, R. & Parisi, D.:In What Kinds of Social Groups can Altruistic Behaviour Evolve? In Conte, R., Hegselmann, R. and Terna, P., Eds., *Simulating Social Phenomena – LNEMS 456, Springer-Verlag, Berlin.*(1997).
20. Hales, *D.:Tag Based Cooperation in Artificial Societies.* Unpublished Ph.D. Thesis, Department of Computer Science, University of Essex (2001).
21. Binmore, K.: *Game Theory and the Social Contract. Volume 2: Just Playing*. The MIT Press, Cambridge, MA. (1998).

Norms and Their Role in a Model of Electronic Institution

Ioan Alfred Letia[1] and Wamberto W. Vasconcelos[2]

[1] Technical University of Cluj-Napoca
Department of Computer Science
Baritiu 28, RO-3400 Cluj-Napoca, Romania
letia@cs.utcluj.ro
http://cs-gw.utcluj.ro/~letia

[2] University of Aberdeen
Department of Computing Science
AB24 3UE Aberdeen, UK
wvasconcelos@acm.org
http://www.csd.abdn.ac.uk/~wvasconc

Abstract. It is convenient for agents participating in computational institutions to have as much autonomy as possible. The model of electronic institution considered in this paper imposes restrictions on the communication between agents through a precise protocol. Otherwise the agents are reasonable autonomous. To foster some kind of acceptable social behaviour norms are used. When conflicting goals appear in the institution, their preference is expressed by norms that show the reward/punishment applied to agents obeying/disobeying a given norm. Several types of agents are defined: social, rebellious, selfish. Their influence on the performance of the institution is captured by utility measures expressing social contribution and individual satisfaction.

1 Introduction

A fundamental challenge of artificial agents is the development of an infrastructure that gives them autonomy to achieve their own goals, while at the same time imposing some kind of social order with its enforcing mechanisms [3]. Specifications for computational societies that take into account agent heterogeneity, conflicting individual goals and limited trust [1] are therefore very significant if artificial multi-agent systems are to be deployed in an organisation. Another approach used to verify behaviour is the checking [14] of the dynamical properties in an organisation.

Autonomous artificial agents have been specified in various guises, some of which have also been developed with a social dimension. A BDI interpreter takes norms and obligations into account in an agent's deliberation [4]. The effects of the environmental factors on the agents' level of social consciousness [11] has also been studied. A conceptual agent model that takes into account its environment by using an action theory, based on the Situation Calculus, has recently proved

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 240–258, 2004.

quite successful as a lightweight tool [10,18]. Knowledge-based programs with sensing in the situation calculus are very promising for the on-line execution of action sequences [19].

Among many ways for viewing sociability, social commitments have been defined through normative policies [17,22]. A very rich body of literature has studied the practical applicability of a combination of deontic logic and a logic of action/agency to a set of regulations [21]. Yet the end result of norms and regulations is their social contribution and individual satisfaction, which has been studied recently in a computational model that constrains autonomy through norms [15].

An advanced model of computational institution that we are aware of is the electronic institution [7,8,20], with a formalisation of the language for institutions and norms [7] and lately an editor [6]. This is an architecturally-neutral e-institution with no assumption on the particular architecture used to develop the agents. The representation we have chosen for the specification of agents is high-level, and therefore it can also be regarded as architecturally-neutral.

In this paper we enhance the skeleton-based development for an electronic institution [23] toward attaining the goal of a working artificial institution where agents can improve in time, thereby achieving a kind of life-cycle [24]. The logic-based specification of this model of e-institution in the next section is intended to give a precise definition of our current understanding of what such an institution can offer for a social environment. We then present a very simple institution world, convenient for the purpose of our study, that is simple enough to show some basic aspects that one would expect in an artificial society [15]. The institutionalized agents that populate this world are described in the Fluent Calculus, a Situation Calculus based formalism. They are meant to give a convenient normative specification; hence we do not commit ourselves to a specific architecture. After presenting some normative positions, that have to be tackled in such an e-institution endowed with a set of rules, we end with a discussion on our current view regarding a computational society that might be deployed in the near future.

2 Logic-Based Electronic Institutions

In the same way that social institutions are somehow forged (say, in print or by common knowledge), the laws that should govern the interactions among heterogeneous agents can be defined by means of electronic institutions (e-institutions, for short). E-institutions are non-deterministic finite-state machines (NDFSM) describing possible interactions among agents [7,8,20]. The interactions are only by means of message exchanges, that is, messages that are sent and received by agents. E-institutions define communication protocols among agents with a view to achieving global and individual goals. We have investigated a logic-based rendition to E-Institutions [25] through which important assumptions are made explicit and explain these below.

Scenes are the basic components of an e-institution: they describe interactions among agents. Scenes describe a very specific scenario of interactions, for example, a scene where agents advertise their goods to sell and receive offers. We define a scene as below:

Def. (Scene) A scene is $\mathbf{S} = \langle R, W, w_0, W_f, WA, WE, f^{Guard}, Edges, f^{Label}\rangle$ where

- $R = \{r_1, \ldots, r_n\}$ is the set of *roles;*
- $W = \{w_0, \ldots, w_m\}$ is a finite, non-empty set of states;
- $w_0 \in W$ is the *initial state*;
- $W_f \subseteq W$ is the non-empty set of *final states;*
- *WA* is a set of sets $WA = \{WA_r \subseteq W \mid r \in R\}$ where each WA_r, $r \in R$, is the set of *access states* for role r;
- *WE* is a set of sets $WE = \{WE_r \subseteq W \mid r \in R\}$ where each WE_r, $r \in R$, is the set of *exit states* for role r;
- $f^{Guard} : WA_r \mapsto \mathcal{L}$ and $f^{Guard} : WE_r \mapsto \mathcal{L}$ associates with each access state WA_r and exit state WE_r of role r a construct of $\mathcal{L}$ described below.
- $Edges \subseteq W \times W$ is a set of *directed edges;*
- $f^{Label} : Edges \mapsto \mathcal{L}$ associates each element of *Edges* with a construct of $\mathcal{L}$.

This definition is a variation of that found in [23]. We have added to access and exit states, via function f^{Guard}, explicit restrictions formulated as formulae of a purpose-built first-order logic described below. The labelling function f^{Label} is defined similarly, but mapping *edges* to our logic formulae, described below.

Within an e-institution sets will be represented by words starting with capital letters and in this font, as in, for example "S", "Set" and "Buyers". Variables will be denoted by words starting with capital letters in *this typefont,* as in, for example, *"X"*, *"Var"* and *"Buyer"*. We shall represent constants by words starting with non-capital letters in *this font;* some examples are *"a"* and *"item"*. We shall assume the existence of a recursively enumerable set *Vars* of variables and a recursively enumerable set *Consts* of constants.

The scenes, as formalised above, are where the communication among agents actually take place. However, individual scenes can be part of a more complex context in which specific sequences of scenes have to be followed. For example, in some kinds of electronic markets, a scene where agents meet other agents to choose their partners to trade is followed by a scene where the negotiations actually take place. We define *transitions* as a means to connect and relate scenes:

Def. (Transition) A transition is the tuple $\mathbf{T} = \langle CI, w_a, \mathcal{L}, w_e, CO\rangle$ where

- $CI \subseteq \bigcup_{i=1}^{n}(WE_i \times w_a)$, is the set of *connections into* the transition, $WE_i, 1 \leq i \leq n$ being the sets of exit states for all roles from all scenes;
- w_a is the *access state* of the transition;
- w_e is the *exit state* of the transition;
- $\mathcal{L}$, a formula of our logic defined below, labels the pair $(w_a, w_e) \mapsto \mathcal{L}$;
- $CO \subseteq \bigcup_{j=1}^{m}(w_e \times WA_j)$, is the set of *connections out of* the transition, $WA_j, 1 \leq j \leq m$ being the sets of access states for all roles onto all scenes.

A transition has only two states w_a, its access state, and w_e, its exit state, and a set of connections *CI* relating the exit states of scenes to w_a and a set of connections *CO* relating w_e to the access states of scenes. The conditions under which agents are allowed to move from w_a to w_e are specified by a formula $\mathcal{L}$ of our set-based logic, introduced below.

Transitions can be seen as simplified scenes where agents' movements can be grouped together and synchronised out of a scene and into another one. The roles of agents may change, as they go through a transition. An important feature of transitions lies in the kinds of formula $\mathcal{L}$ we are allowed to use. Contrary to scenes, where there can only be references to constructs within the scene, within a transition we can make references to constructs of any scene that connects to the transition. This difference is formally represented when we define the semantics of our constructs, below.

Our e-institutions are collections of scenes and transitions appropriately designed. Formally:

Def. (E-Institution) An e-institution is $\mathbf{E} = \langle \mathit{Scenes}, \mathbf{S}_0, \mathbf{S}_f, \mathit{Trans}\rangle$ where

- $\mathit{Scenes} = \{\mathbf{S}_0, \ldots, \mathbf{S}_n\}$ is a finite and non-empty set of *scenes;*
- $\mathbf{S}_0 \in \mathit{Scenes}$ is the *root scene*;
- $\mathbf{S}_f \in \mathit{Scenes}$ is the *output* scene;
- $\mathit{Trans} = \{\mathbf{T}_0, \ldots, \mathbf{T}_m\}$ is a finite and non-empty set of *transitions;*

We shall impose the restriction that the transitions of an e-institution can only connect scenes from the set *Scenes,* that is, for all $\mathbf{T} \in \mathit{Trans}$, $\mathit{CI} \subseteq \bigcup_{i=0}^{n}(\mathit{WE}_i \times w_a), i \neq f$ (the exit states of the output scene can not be connected to a transition) and $\mathit{CO} \subseteq \bigcup_{j=1}^{n}(w_e \times \mathit{WA}_j)$ (the access state of the root scene cannot be connected to a transition).

For the sake of simplicity, we have not included in our definition above the *normative rules* [7] which capture the obligations agents get bound to as they exchange messages. We are aware that this makes our definition above closer to the notion of *performative structure* [7] rather than an e-institution.

2.1 $\mathcal{L}$: A Set-Based Logic for E-institutions

In this section, we describe a set-based first-order logic $\mathcal{L}$ employed in the above definitions. The logic $\mathcal{L}$ provides us with a compact notation with which we can formally describe conditions on scenes, transitions and, ultimately, e-institutions. Intuitively, these constructs define (pre- and post-) conditions that should hold in order for agents to move through an e-institution. We define $\mathcal{L}$ as below:

Def. (Syntax of $\mathcal{L}$) $\mathcal{L}$ consists of formulae $\mathit{Qtf}\ (\mathit{Atfs} \Rightarrow \mathit{SetCtrs})$ where Qtf is the *quantification,* Atfs is a conjunction of *atomic formulae* and $\mathit{SetCtrs}$ is a conjunction of *set constraints,* as explained below.

Qtf provides our constructs with universal and existential quantification over (finite) sets; Atfs expresses atomic formulae that must hold true and $\mathit{SetCtrs}$ represents set constraints that (are made to) hold true. We define the classes of constructs Qtf, Atfs and $\mathit{SetCtrs}$ in the sequel. In order to define the class Atfs of atomic formulae conjunctions, we first put forth the concept of *terms:*

Def. (Terms) All elements from *Vars* and *Consts* are in *Terms,* If $t_1, \ldots, t_n$ are in *Terms,* then $f(t_1, \ldots, t_n)$ is also in *Terms,* where f is any functional symbol.

The class of terms *Terms* is thus defined recursively, based on variables and constants and their combination with functional symbols. Recall our typographic conventions for variables and constants above. We can now define the class *Atfs* of conjunctions of atomic formulae:

Def. (Conjunctions of Atomic Formulae) If $t_1, \ldots, t_n$ are Terms, then $p(t_1, \ldots, t_n)$ is an atomic formula (or, simply, an *Atf*) of *Atfs* , where p is any predicate symbol. A special atomic formula is defined via the "=" symbol, as $t_1 = t_2$. Furthermore, for any Atf_1 and Atf_2 in *Atfs*, the construct $Atf_1 \wedge Atf_2$ is also in *Atfs*.

This is another recursive definition: the basic components are the simple atomic formulae built with terms. These components (and their combinations) can be put together as conjuncts. We now define the class of *set constraints.* These are restrictions on set operations such as union, intersection, cartesian product and set difference, following their usual definition [12]:

Def. (Set Constraints) The set constraints are a conjunction of set operations. The allowed set constraints are defined by the following grammar:

$$\begin{array}{rcl}
SetCtrs & \rightarrow & SetCtrs \wedge SetCtrs \mid (SetCtrs) \mid MTest \mid SetProp \\
MTest & \rightarrow & Term \in SetOp \mid Term \notin SetOp \\
SetProp & \rightarrow & card(SetOp)\ Op\ \mathbb{N} \mid card(SetOp)\ Op\ card(SetOp) \mid SetOp = SetOp \\
Op & \rightarrow & = \mid > \mid \geq \mid < \mid \leq \\
SetOp & \rightarrow & SetOp \cup SetOp \mid SetOp \cap SetOp \mid SetOp - SetOp \mid SetOp \times SetOp \mid \\
 & & (SetOp) \mid \mathsf{Set} \mid \emptyset
\end{array}$$

MTest is a *membership test,* that is, a test whether an element belongs or not to the result of a set operation *SetOp* (in particular, to a specific set). *SetProp* represents the *set properties,* that is, restrictions on set operations as regards to their size *(card)* or their contents. ℕ is the set of natural numbers. *Op* stands for the allowed operators of the set properties. *SetOp* stands for the *set operations,* that is, expressions whose final result is a set. An example of a set constraint is $B \in \mathsf{Buyers} \wedge card(\mathsf{Buyers}) \geq 0 \wedge card(\mathsf{Buyers}) \leq 10$. We may, alternatively, employ $|\mathsf{Set}|$ to refer to the cardinality of a set, that is, $|\mathsf{Set}| = card(\mathsf{Set})$.Additionally, in order to simplify our set expressions and improve their presentation, we can use $0 \leq |\mathsf{Buyers}| \leq 10$ instead of the expression above.

Finally, we define the quantifications *Qtf*:

Def. (Quantifications) The quantification *Qtf* is defined via the following grammar:

$$\begin{array}{rcl}
Qtf & \rightarrow & Qtf'\ Qtf \mid Qtf' \\
Qtf' & \rightarrow & Q\ Var \in SetOp \mid Q\ Var \in SetOp,\ Var = Term \\
Q & \rightarrow & \forall \mid \exists \mid !\exists
\end{array}$$

Where *SetOp* is any set operation, *Term* is any term (see above) and *Var* is any variable from *Vars*.

We pose an important additional restriction on our quantifications: either *Var* or subterms of *Term* must occur in $(Atfs \Rightarrow SetCtrs)$.

Using the typographic conventions presented above, we can now build correct formulae as they appear in our e-institutions. An example of such construct, is $\exists B \in \mathsf{Ags}\,(m(B, adm, enter(buyer)) \Rightarrow (B \in \mathsf{Bs} \wedge 1 \leq |\mathsf{Bs}| \leq 10))$. To simplify our formulae, we shall also write quantifications of the form $Qtf\ Var \in SetOp,\ Var = Term$ simply as $Qtf\ Term \in SetOp$. For instance, $\forall X \in \mathsf{Set}, X = f(a, Z)$ will be written as $\forall f(a, Z) \in \mathsf{Set}$.

2.2 The Semantics of $\mathcal{L}$

In this section we show how the formulae of $\mathcal{L}$ are mapped to truth values $\top$ (true) or $\bot$ (false). For that, we define first the *interpretation* for $\mathcal{L}$ formulae:

Def. (Interpretation $\Im$) An interpretation $\Im$ for a formula of $\mathcal{L}$ is the pair $\Im = (\sigma, \Omega)$ where σ is a possibly empty set of ground atomic formulae (*i.e.* atfs without variables) and Ω is a set of sets.

Intuitively our interpretations provide in σ what is required to determine the truth value of $Qtf(Atfs)$ and in Ω what is needed in order to assign a truth value to $Qtf(SetCtrs)$.

We did not include in our definition of interpretation above the notion of *universe of discourse* (also called *domain*) nor the usual mapping between constants and elements of this universe, neither the mapping between function and predicate symbols of the formula and functions and relations in the universe of discourse [5,16]. This is because we are only interested in the relationships between $Atfs$ and $SetCtrs$ and how we can automatically obtain an interpretation for a given formula. However, we can define the union of all sets in Ω as our domain. It is worth mentioning that the use of a set of sets to represent Ω does not cause undesirable paradoxes: since we do not allow the formulae in $\mathcal{L}$ to make references to Ω, but only to sets in Ω, this will not happen.

We define the semantic mapping $\mathbf{k} : \mathcal{L} \times \Im \mapsto \{\top, \bot\}$ as follows:

1. $\mathbf{k}(\forall Terms \in SetOp\ \mathcal{L}, \Im) = \top$ iff $\mathbf{k}(\mathcal{L}|_e^{Terms}, \Im) = \top$ for all $e \in \mathbf{k}'(SetOp, \Im)$.
   $\mathbf{k}(\exists Terms \in SetOp\ \mathcal{L}, \Im) = \top$ iff $\mathbf{k}(\mathcal{L}|_e^{Terms}, \Im) = \top$ for some $e \in \mathbf{k}'(SetOp, \Im)$.
   $\mathbf{k}(\exists! Terms \in SetOp\ \mathcal{L}, \Im) = \top$ iff $\mathbf{k}(\mathcal{L}|_e^{Terms}, \Im) = \top$ for a single $e \in \mathbf{k}'(SetOp, \Im)$.
2. $\mathbf{k}((Atfs \Rightarrow SetCtrs), \Im) = \bot$ iff $\mathbf{k}(Atfs, \Im) = \top$ and $\mathbf{k}(SetCtrs, \Im) = \bot$.
3. $\mathbf{k}(Atfs_1 \wedge Atfs_2, \Im) = \top$ iff $\mathbf{k}(Atfs_1, \Im) = \mathbf{k}(Atfs_2, \Im) = \top$.
   $\mathbf{k}(Atf, \Im) = \top$ iff $Atf \in \sigma, \Im = (\sigma, \Omega)$.
4. $\mathbf{k}(SetCtrs_1 \wedge SetCtrs_2, \Im) = \top$ iff $\mathbf{k}(SetCtrs_1, \Im) = \mathbf{k}(SetCtrs_2, \Im) = \top$.
   $\mathbf{k}(Terms \in SetOp, \Im) = \top$ iff $Terms \in \mathbf{k}'(SetOp, \Im)$;
   $\mathbf{k}(Terms \not\in \mathcal{S}\rceil\sqcup\mathcal{O}_{\surd}, \Im) = \top$ iff $Terms \not\in \mathbf{k}'(SetOp, \Im)$
5. $\mathbf{k}(|SetOp|\ Op\ \mathbb{N}, \Im) = \top$ iff $|\mathbf{k}'(SetOp, \Im)|\ Op\ \mathbb{N}$ holds.
   $\mathbf{k}(|SetOp_1|\ Op\ |SetOp_2|, \Im) = \top$ iff $|\mathbf{k}'(SetOp_1, \Im)|\ Op\ |\mathbf{k}'(SetOp_2, \Im)|$ holds.
   $\mathbf{k}(SetOp_1 = SetOp_2, \Im) = \top$ iff $\mathbf{k}'(SetOp_1, \Im) = \mathbf{k}'(SetOp_2, \Im)$.

In item 1 we address the three quantifiers over $\mathcal{L}$ formulae, where $\mathcal{L}|_e^{Terms}$ is the result of replacing every occurrence of *Terms* by e in $\mathcal{L}$. Item 2 describes the usual meaning of the right implication. Item 3 formalises the meaning of conjunctions $Atfs$ and the basic case for individual atomic formulae – these are only considered true if they belong to the associated set σ of the interpretation $\Im$. Item 4 formalises the meaning of the conjunct and disjunct operations over set constraints $SetCtrs$ and the basic membership test to the result of a set operation $SetOp$. Item 5 describes the truth-value of the distinct set properties $SetProp$. These definitions describe only one case of the mapping: since ours is a total mapping, the situations which are not described represent a mapping with the remaining value $\top$ or $\bot$.

The auxiliary mapping $\mathbf{k}' : SetOp \times \Im \mapsto \mathsf{Set}$ in $\Omega, \Im = (\sigma, \Omega)$, referred to above and which gives meaning to the set operations is thus defined:

- $\mathbf{k}'(SetOp_1 \cup SetOp_2, \Im) = \{e \mid e \in \mathbf{k}'(SetOp_1, \Im) \text{ or } e \in \mathbf{k}'(SetOp_2, \Im)\}$
- $\mathbf{k}'(SetOp_1 \cap SetOp_2, \Im) = \{e \mid e \in \mathbf{k}'(SetOp_1, \Im) \text{ and } e \in \mathbf{k}'(SetOp_2, \Im)\}$
- $\mathbf{k}'(SetOp_1 - SetOp_2, \Im) = \{e \mid e \in \mathbf{k}'(SetOp_1, \Im) \text{ and } e \not\in \mathbf{k}'(SetOp_2, \Im)\}$
- $\mathbf{k}'(SetOp_1 \times SetOp_2, \Im) = \{(e_1, e_2) \mid e_1 \in \mathbf{k}'(SetOp_1, \Im) \text{ and } e_2 \in \mathbf{k}'(SetOp_2, \Im)\}$
- $\mathbf{k}'((SetOp), \Im) = (\mathbf{k}'(SetOp, \Im))$.
- $\mathbf{k}'(\mathsf{Set}, \Im) = \{e \mid e \in \mathsf{Set} \text{ in } \Omega, \Im = (\sigma, \Omega)\}$, $\mathbf{k}'(\emptyset, \Im) = \emptyset$.

The four set operations are respectively given their usual definitions. The meaning of a particular set Set is its actual contents, as given by Ω in $\Im$. Lastly, the meaning of an empty set $\emptyset$ in a set operation is, of course, the empty set.

We are interested in *models* for our formulae, that is, interpretations that map $\mathcal{L}$ to the truth value $\top$ (true). We are only interested in those interpretations in which *both* sides of the "$\Rightarrow$" in the $\mathcal{L}$'s hold true. Formally:

Def. (Models) An interpretation $\Im$ is a *model* for $\mathcal{L} = Qtf\ (Atfs \Rightarrow SetCtrs)$, denoted by $\mathbf{m}(\mathcal{L}, \Im)$ iff $\mathbf{k}(Qtf\ Atfs, \Im) = \mathbf{k}(Qtf\ SetCtrs, \Im) = \top$.

The scenarios arising when the left-hand side of the $\mathcal{L}$'s is false do not interest us: we want this formalisation to restrict the meanings of our constructs only to those desirable (correct) ones. The study of the anomalies and implications caused by not respecting the restrictions of an e-institutions albeit important is not in the scope of this work.

We now define the extension of an interpretation, useful when we want to build models for more than one formula $\mathcal{L}$:

Def. (Extension of Interpretation) $\Im' = (\sigma', \Omega')$ is an extension of $\Im = (\sigma, \Omega)$ which accommodates $\mathcal{L}$, denoted by $\mathbf{ext}(\Im, \mathcal{L}) = \Im'$, iff $\mathbf{m}(\mathcal{L}, \Im'')$, $\Im'' = (\sigma'', \Omega'')$ and $\sigma' = \sigma \cup \sigma''$, $\Omega' = \Omega \cup \Omega''$.

2.3 Models for Logic-Based E-institutions

In this section we introduce models for scenes, transitions and e-institutions using the definitions above. A model for a scene is a gradual extension of a model to accommodate all the formulae $\mathcal{L}$'s connecting the initial state to one of the final states. More formally:

Def. (Models of Scenes) An interpretation $\Im$ is a model for a scene $\mathbf{S} = \langle R, W, w_0, W_f, WA, WE, f^{Guard}, Edges, f^{Label}\rangle$, given an initial interpretation $\Im_0$, denoted by $\mathbf{m}(\mathbf{S}, \Im)$, iff $\Im = \Im_n$, where:

- $f^{Label}(w_{i-1}, w_i) = \mathcal{L}_i, 1 \leq i \leq n, w_n \in W_f$, are the formulae labelling edges which connect the initial state w_0 to a final state w_n.
- for $w_i \in WA_r$ or $w_i \in WE_r$ for some role r, that is, w_i is an access or exit state, then $f^{Guard}(w_i) = \mathcal{L}_i^{WA}$ or $f^{Guard}(w_i) = \mathcal{L}_i^{WE}$, respectively.
- for $1 \leq i \leq n, \Im_i = \begin{cases} \mathbf{ext}(\mathbf{ext}(\Im_{i-1}, \mathcal{L}_i^{WA}), \mathcal{L}_i), \text{ if } w_i \in WA_r \\ \mathbf{ext}(\mathbf{ext}(\Im_{i-1}, \mathcal{L}_i^{WE}), \mathcal{L}_i), \text{ if } w_i \in WE_r \\ \mathbf{ext}(\Im_{i-1}, \mathcal{L}_i), \qquad \text{otherwise} \end{cases}$

A model for a scene is built using the formulae that label edges connecting the initial state to a final state. The formulae guarding access and exit states are also taken into account: they are used to extend the model of the previous formulae and this extension is further employed with the formula connecting the state onwards. Since there might be more than one final state and more than one possible way of going from the initial state to a final state, models for scenes are not unique. One should also notice that the existential quantification allows for the *choice* of components for the sets in Ω and hence more potential for different models. In order to obtain a model for a scene, an initial model $\Im_0$, possibly empty, must be provided.

The model of a transition extends the models of scenes connecting with it:

Def. (Models of Transitions) An interpretation $\Im$ is a model for a transition $\mathbf{T} = \langle CI, w_a, \mathcal{L}, w_e, CO \rangle$, denoted by $\mathbf{m}(\mathbf{T}, \Im)$,iff

- $\mathbf{S}_1, \ldots, \mathbf{S}_n$ are all the scenes that connect with *CI, i.e.* the set WE_i of exit states of each scene $\mathbf{S}_i, 1 \leq i \leq n$, has at least one element $WE_{i,r} \times w_a$ in *CI*, and
- $\mathbf{m}(\mathbf{S}_i, \Im_i), \Im_i = (\sigma_i, \Omega_i), \Im' = (\bigcup_{i=1}^{n} \sigma_i, \bigcup_{i=1}^{n} \Omega_i), 1 \leq i \leq n$, and $\mathbf{ext}(\Im', \mathcal{L}) = \Im$

The model of a transition is an extension of the union of the models of all its connecting scenes to accommodate $\mathcal{L}$. Finally, we define the meaning of e-institutions:

Def. (Models of E-Institutions) An interpretation $\Im$ is a model for an e-institution $\mathbf{E} = \langle \mathit{Scenes}, \mathbf{S}_0, \mathbf{S}_f, \mathit{Trans} \rangle$, denoted by $\mathbf{m}(\mathbf{E}, \Im)$,iff

- $\mathit{Scenes} = \{\mathbf{S}_0, \ldots, \mathbf{S}_n\}, \mathbf{m}(\mathbf{S}_i, \Im), 0 \leq i \leq n$; and
- $\mathit{Trans} = \{\mathbf{T}_0, \ldots, \mathbf{T}_m\}, \mathbf{m}(\mathbf{T}_j, \Im), 0 \leq j \leq m$.

2.4 Automatically Building Models

Building a model $\Im$ is a computationally expensive task, involving combinatorial efforts to find the atomic formulae that ought to be in σ and the contents of the sets in Ω. If, however, the formulae $\mathcal{L}$'s of a scene have a simple property, *viz.* the quantification of each formula $\mathcal{L}_i$ only refers to sets that appear on preceding formulae $\mathcal{L}_j, j < i$, then we can build an interpretation gradually, taking into account each formula at a time. This property can be syntactically checked: we can ensure that all sets appearing in $\mathcal{L}_i$'s quantification appears on the right-hand side of a $\mathcal{L}_j$ which leads on to $\mathcal{L}_i$ in a scene. Only if all scenes and transitions of an e-institution fulfil this property is that we can automatically build a model for it.

Assuming this property holds in our e-institutions, then we can build for any formula $\mathcal{L}_i$ a model $\Im_i$ that uses the $\Im_{i-1}$ of the preceding formula. The models of a scene are then built gradually, each formula at a time, via $\mathbf{ext}(\Im_{i-1}, \mathcal{L}_i) = \Im_i$. We assume an initial interpretation $\Im = (\emptyset, \Omega)$ in which Ω is possibly empty or may contain any initial values of sets, so that we can start building the models of the ensuing formulae.

Given $\Im_{i-1}$ and $\mathcal{L}_i$ we can automatically compute $\mathbf{ext}(\Im_{i-1}, \mathcal{L}_i) = \Im_i$. Since the quantifiers of $\mathcal{L}_i$ only refer to sets of the right-hand side of preceding $\mathcal{L}_j$, then $\Im_{i-1}$ should have the actual contents of these sets. We exhaustively generate values for the quantified variables – this is only possible because all the sets are finite – and hence we can assemble the atomic formulae for a possible σ_i of $\Im_i$. With this σ and Ω_{i-1} we then assemble Ω_i, an extension of Ω_{i-1} which satisfies the set constraints of $\mathcal{L}_i$.

2.5 An Example: An Agora Room

To illustrate the definitions above, we provide in Figure 1 a simple example of a scene for an agora room in which agents willing to acquire goods interact with agents intending to sell such goods. This agora scene has been simplified – no auctions or negotiations are contemplated. The sellers announce the goods they want to sell, collect the replies from buyers (all buyers must reply) and confirm the replies. The simplicity of this scene is deliberate, so as to allow us to fully represent and discuss it. A more friendly visual rendition of the formal definition is employed in the figure and is explained below.

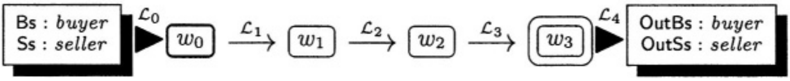


Fig. 1. Diagrammatic Representation for Agora Room Scene

The states $W = \{w_0, w_1, w_2, w_3\}$ are displayed in oval boxes and $Edges = \{(w_0, w_1), (w_1, w_2), (w_2, w_3)\}$ are shown as arrows: if $(w_i, w_j) \in Edges$, then $w_i \longrightarrow w_j$. The initial state w_0 is shown enclosed in a thicker oval box; the final state $W_f = \{w_3\}$ is enclosed in a double oval box. We define the set of roles as $R = \{seller, buyer\}$. An access state $w \in WA$ is marked with a "▶" pointing towards the state with a box containing the role(s) of the agents that may enter the scene at that point and a set name. Exit states are also marked with a "▶" but pointing away from the state; they are also shown with a box containing the roles of the agents that may leave the scene at that point and a set name. We have defined the formulae $\mathcal{L}_i, 0 \leq i \leq 4$, as:

$$\mathcal{L}_0\text{: } \exists B, S \in \mathsf{Ags}\left(\begin{pmatrix} m(B, adm, enter(buyer)) \wedge \\ m(S, adm, enter(seller)) \end{pmatrix} \Rightarrow \begin{pmatrix} B \in \mathsf{Bs} \wedge 1 \leq |\mathsf{Bs}| \leq 10 \wedge \\ S \in \mathsf{Ss} \wedge 1 \leq |\mathsf{Ss}| \leq 10 \end{pmatrix}\right)$$

$$\mathcal{L}_1\text{: } \forall S \in \mathsf{Ss}\, \forall B \in \mathsf{Bs}\, \exists I \in \mathsf{Is}\, (m(S, B, offer(I)) \Rightarrow \langle S, B, I \rangle \in \mathsf{Ofs})$$

$$\mathcal{L}_2\text{: } \forall \langle S, B, I \rangle \in \mathsf{Ofs}\, \exists! A \in \mathsf{Ans}\, (m(B, S, reply(I, A)) \Rightarrow \langle B, S, I, A \rangle \in \mathsf{Prs})$$

$$\mathcal{L}_3\text{: } \forall \langle B, S, I, ok \rangle \in \mathsf{Prs}\, \exists! A \in \mathsf{Ans}\, (m(S, B, confirm(I, A)) \Rightarrow \langle S, B, I, A \rangle \in \mathsf{Rs})$$

$$\mathcal{L}_4\text{: } \forall B \in \mathsf{Bs}\, \forall S \in \mathsf{Ss}\left(\begin{pmatrix} m(B, adm, leave) \wedge \\ m(S, adm, leave) \end{pmatrix} \Rightarrow \begin{pmatrix} B \in \mathsf{OutBs} \wedge S \in \mathsf{OutSs} \wedge \\ \mathsf{OutBs} = \mathsf{Bs} \wedge \mathsf{OutSs} = \mathsf{Ss} \end{pmatrix}\right)$$

The left-hand side of the $\mathcal{L}_i$ are atomic formulae which must hold in σ_i and the right-hand side are set constraints that must hold in Ω_i. The atomic formula stand for messages exchanged among the agents as they move along the edges of the scene. The above definitions give rise to the following semantics:

$$\begin{aligned}
&\mathbf{ext}\left(\left(\emptyset, \begin{Bmatrix}\mathsf{Ags} = \{ag_1, \ldots, ag_4\}, \\ \mathsf{Ans} = \{ok, not\_ok\}, \\ \mathsf{Is} = \{car, boat, plane\}\end{Bmatrix}\right), \mathcal{L}_0\right) = \left(\begin{Bmatrix}m(ag_1, adm, enter(seller)) \\ m(ag_2, adm, enter(buyer)) \\ m(ag_3, adm, enter(buyer))\end{Bmatrix}, \begin{Bmatrix}\mathsf{Ags}, \mathsf{Ans}, \mathsf{Is}, \\ \mathsf{Bs} = \{ag_2, ag_3\} \\ \mathsf{Ss} = \{ag_1\}\end{Bmatrix}\right) = \\
&\Im_0 \\
&\mathbf{ext}(\Im_0, \mathcal{L}_1) = \left(\begin{matrix}\sigma_0 \cup \\ \begin{Bmatrix}m(ag_1, ag_2, offer(car)) \\ m(ag_1, ag_3, offer(boat))\end{Bmatrix}\end{matrix}, \begin{Bmatrix}\mathsf{Ags}, \mathsf{Ans}, \mathsf{Is}, \mathsf{Bs}, \mathsf{Ss}, \\ \mathsf{Ofs} = \begin{Bmatrix}\langle ag_1, ag_2, car\rangle, \\ \langle ag_1, ag_3, boat\rangle\end{Bmatrix}\end{Bmatrix}\right) = \Im_1 \\
&\mathbf{ext}(\Im_1, \mathcal{L}_2) = \left(\begin{matrix}\sigma_1 \cup \\ \begin{Bmatrix}m(ag_2, ag_1, reply(car, ok)) \\ m(ag_3, ag_1, offer(boat, not\_ok))\end{Bmatrix}\end{matrix}, \begin{Bmatrix}\mathsf{Ags}, \mathsf{Ans}, \mathsf{Is}, \mathsf{Bs}, \mathsf{Ss}, \mathsf{Ofs}, \\ \mathsf{Prs} = \begin{Bmatrix}\langle ag_2, ag_1, car, ok\rangle, \\ \langle ag_3, ag_1, boat, not\_ok\rangle\end{Bmatrix}\end{Bmatrix}\right) = \\
&\Im_2 \\
&\mathbf{ext}(\Im_2, \mathcal{L}_3) = \left(\begin{matrix}\sigma_2 \cup \\ \{m(ag_1, ag_2, confirm(car, not\_ok))\}\end{matrix}, \begin{Bmatrix}\mathsf{Ags}, \mathsf{Ans}, \mathsf{Is}, \mathsf{Bs}, \mathsf{Ss}, \mathsf{Ofs}, \mathsf{Prs}, \\ \mathsf{Rs} = \{\langle ag_1, ag_2, car, not\_ok\rangle\}\end{Bmatrix}\right) = \Im_3 \\
&\mathbf{ext}(\Im_3, \mathcal{L}_4) = \left(\begin{matrix}\sigma_3 \cup \\ \begin{Bmatrix}m(ag_1, adm, leave)) \\ m(ag_2, adm, leave)) \\ m(ag_3, adm, leave))\end{Bmatrix}\end{matrix}, \begin{Bmatrix}\mathsf{Ags}, \mathsf{Ans}, \mathsf{Is}, \mathsf{Bs}, \mathsf{Ss}, \mathsf{Ofs}, \mathsf{Prs}, \mathsf{Rs}, \\ \mathsf{OutBs} = \{ag_2, ag_3\} \\ \mathsf{OutSs} = \{ag_1\}\end{Bmatrix}\right) = \Im_4
\end{aligned}$$

2.6 Enacting Logic-Based E-institutions

We have incorporated the concepts above into a distributed enactment platform. This platform, implemented in SICStus Prolog [13], uses the semantics of our constructs to perform a simulation of an e-institution. The platform relies on a number of administrative agents, implemented as independent processes, to overlook the enactment, building models and interacting with the agents partaking the enactment via a blackboard architecture, using SICStus Linda tuple space [2,13].

The platform starts up for each scene an administrative agent *sceneAdm*. An initial model is available for all scenes, $\Im = (\emptyset, \Omega)$ where Ω (possibly empty) contains the values of any sets that need to be initially defined. Some of such sets are, for instance, the identity of those agents that may join the e-institution, the possible values for items and their prices, and so on. Agent *sceneAdm* follows the edges of a scene, starting from w_0 and, using $\Im$, creates the set σ_0 of atomic formulae. The set σ_0 is assembled by evaluating the quantification of $\mathcal{L}_0$ over the sets in Ω.

An enactment of an e-institution begins with the enactment of the root scene and terminates when all agents leave the output scene. Engineers may specify whether a scene can have many instances enacted simultaneously, depending on the number and order of agents willing to enter it. We did not include this feature in our formal presentation because in logic-theoretic terms instances of a scene can be safely seen as different scenes: they are enacted independently from each other, although they all conform to the same specification.

Our platform takes into account the agents that will partake in it. These are called the *performing agents* and are automatically synthesised from the description of the e-institution, as described in [23]. A performing agent sends a message by checking if the corresponding σ set contains the message it wants to send; if the message is available then the agent "sends" it by marking it as sent. This mark is for the benefit of the *admScene* agent: the *admScene* agent creates *all* messages that can be sent, but not all of them may in fact be sent.

The messages that have been marked as sent are those that were actually sent by the performing agents.

Similarly, a performing agent receives a message by marking it as received. However, it can only receive a message that has been previously marked as sent by another agent. Both the sending and receiving agents use the format of the messages to ensure they conform to the format specified in the edge they are following. To ensure that an agent does not try to receive a message that has not yet been marked as sent but that may still be sent by some agent, the *admScene* agent synchronises the agents in the scene: it first lets the sending agents change state by moving along the corresponding edge, marking their messages as sent. When all sending agents have moved, then the *admScene* agent lets the receiving agents receive their messages and move to the following state of the scene.

The synchronisation among the agents of a scene is achieved via a simple semaphore represented as a term in the tuple space. The performing agents trying to send a message must wait until this semaphore has a specific value. Likewise, the agents that will receive messages are locked until the semaphore allows them to move. The performing agents inform to the *admScene* agent, via the tuple space, the state of the scene they are currently at. With this information the *admScene* agent is able to "herd" agents from one state to another, as it creates messages templates, lets the sending agents mark them as sent and then lets the receiving agents mark them as received (also retrieving their contents). Those agents that do not send nor receive can move between states without having to wait for the semaphore. All agents though synchronise at every state of the scene, that is, there is a moment in the enactment when all agents are at state w_i, then after sending and receiving (or just moving) they are all at state w_{i+1}.

Transitions are enacted in a similar fashion. The platform assigns an agent *admTrans* to look after each transition. Transitions, however, differ from scenes in two ways. Firstly, we do not allow instances of transitions. This is strictly a methodological restriction, rather than a technical one: we want transitions to work as "meeting points" for agents moving between scenes and instances of transitions could prevent this. Secondly, transitions are *permanent,* that is, their enactment never comes to an end. Scenes (or their instances), once enacted (*i.e.* all the agents have left it at an exit state), cease to exist, that is, the *admScene* agent looking after it stops.

When a scene comes to an end, the *admScene* agent records in the tuple space the model it built as a result of the scene's enactment. The atomic formulae are only important during the enactment since they actively define the interpretations being built. However, only the sets in the Ω part of the interpretation is left as a record of the enactment. This is useful for following the dynamics of the e-institution, and it is also essential for the transitions. The *admTrans* agents looking after transitions use the sets left behind by the *admScene* agents to build their models.

3 A Sample Normative Institution

The world chosen in this paper as running example is a basic scenario with agents having to perform certain tasks in an abstract organisation environment. Although their interaction is extremely simple in this model, they must consider the interest of the organisation, if the organisation on the whole is to increase its gain. The cooperation of the individual agents is enforced by a reward/punishment scheme, given by very simple norms. We must emphasize that this simplicity has been deliberately chosen to keep the presentation as simple as possible.

3.1 Norms for Agent Tasks

The norms are defined by preferences among the various tasks the agents have to perform within the organisation.

Norms for Tasks

- norm showing preference for task T_k over T_i:
  $$N_{ik} : T_i \prec T_k$$
- norm with utility expressing reward (if positive) and/or punishment (if negative)
  $$N_{ik} : T_i(U_i) \prec T_k(U_k)$$

However, if just preferences are used, with no reward/punishment scheme, the agents will not have to comply with the norms. For this reason we use a reward/punishment scheme, which specifies a punishment in the case an agent does not comply with a given norm, and a reward when it does.

In the current scenario the environment may evolve in the manner shown below.

Trace of Tasks for Institutionalized Agents

Environment Time	Task Set	Norms on Tasks
0	$T_1, T_2, T_4, T_5, T_6, T_7, T_9$	$T_1(3) \prec T_4(8)$, $T_9(-12) \prec T_7(7)$
1	T_1, T_3, T_5, T_8	$T_3(-16) \prec T_5(15)$, $T_8(-10) \prec T_3(20)$

At environment time 0 tasks T_1,T_2,T_4,T_5,T_6,T_7,T_9 are waiting to be served. The assumption is that agent P_1 is a social agent and complies to the active norm $T_1(3) \prec T_4(8)$, while P_9 is rebellious and disregards norm $T_9(-12) \prec T_7(7)$. P_9 will be punished with a decrease in its utility of 12 points, and the same punishment will be applied to the institution, since T_9 should not have been processed at this time, in the presence of T_7. P_3 is a selfish agent and can gain 20 units from the active norm $T_8(-10) \prec T_3(20)$ at time 1, while being punished with just 16 units due to the active norm $T_3(-16) \prec T_5(15)$, therefore gaining an overall 4 units. P_3 is not concerned about the 16 units lost by the group.

3.2 Communication Protocol

To perform experiments in this simple scenario we need to define the communication protocol, shown in Figure 2. We have one administrator and several participants in such a scene, with the communicative acts shown below.

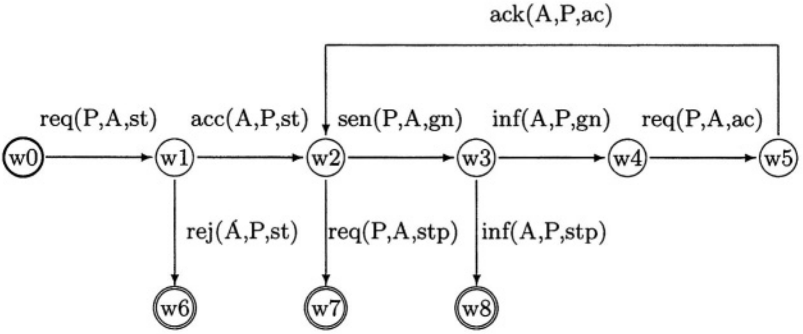


Fig. 2. Administrator-participant communication protocol

Communication Protocol

- $req(P, A, st)$ participant agent requests administrator to start
- $acc(A, P, st)$ administrator accepts participant agent to start
- $rej(A, P, st)$ administrator rejects request of participant agent to start
- $sen(P, A, gn)$ participant agent senses the current goal/norm
- $req(P, A, stp)$ participant agent requests to stop
- $inf(A, P, gn)$ administrator informs participant agent on goal/norm
- $inf(A, P, stp)$ administrator inform participant agent that it must stop
- $ack(A, P, ac)$ administrator acknowledges participant agent about execution of action

A participant can enter the scene by requesting the administrator to start work in the organisation $req(P, A, st)$, and the administrator can accept its participation $acc(A, P, st)$, or reject it $rej(A, P, st)$. When running the experiment, the participant can sense for the current task (goal/norm) $sen(P, A, gn)$, or request to stop activity $req(P, A, stp)$. The administrator informs the participant agent on the task $inf(A, P, gn)$, or that it must stop activity $inf(M, A, stp)$. It also acknowledges the participant agent about the execution of its current action $ack(A, P, ac)$.

3.3 Situation Calculus Specification Language

For the specification of agents populating the institution we use the Golog[1] language [10,18], a version of action theory in the situation calculus. Such a high-level programming language is very convenient for describing the behaviour of agents in dynamic and incompletely known worlds. The main syntactic constructs of the language are depicted in the following definition.

[1] http://www.cs.toronto.edu/ea/~cogrobo/

The Golog language is a high-level programming language with the constructs:

- α is a primitive action
- $\phi?$ is a wait for a condition
- $(\delta_1; \delta_2)$ is a sequence
- $(\delta_1|\delta_2)$ is a nondeterministic choice of action
- $\pi x.\delta$ is a nondeterministic choice of arguments
- δ^* is a nondeterministic iteration
- **if** ϕ **then** δ_1 **else** δ_2 is a synchronized conditional
- **while** ϕ **do** δ is a synchronized loop
- **proc** $\mathrm{P}(\boldsymbol{v})$ δ **end** is a procedure definition.

Apart from primitive actions α and waiting for conditions $\phi?$, the language consists of sequences $(\delta_1; \delta_2)$ and nondeterministic choices $(\delta_1|\delta_2)$. More complex actions can be expressed by the nondeterministic choice of arguments $\pi x.\delta$, or the usual **if, while,** and **proc** constructs.

3.4 Fluents Describing Agent Dynamics

In our reward/punishment scenario, we use a basic action theory for the above primitive actions and fluents. The action theory defines the actions available to an agent and its effects on the environment. It also specifies events that can occur in the environment that the agents inhabit.

Ordinary primitive actions are the actions that agents can execute (their capabilities) in the given environment.

Ordinary Primitive Actions

- *senseTaskNorm* senses the current task and norm (if any)
- *actOnTask* executes action to achieve current task
- *noAction* executes nothing

For instance, our agents can *senseTaskNorm,* that is they can find out whether there is any task to be carried out and its corresponding norm, if there is a norm. At this level of abstraction we have our agents to *actOnTask,* or execute *noAction,* if no task is available for them to act upon.

Exogenous primitive actions are actions that represent events in the environment, that are exogeneous to the agent.

Exogenous Primitive Actions

- $reqTask(T_i)$ a request to achieve task T_i
- $taskUtility(T_i, U_i)$ task T_i has utility U_i
- $conflictingTasks(T_i, T_j)$ tasks T_i, T_j are in conflict
- $T_i(U_i) \prec T_k(U_k)$ task T_k gives utility $U_k > U_i$

In our scenario a $reqTask(T_i)$ might appear, when task T_i is expected to be served by agent P_i. At the same time, a task might have its utility U_i specified by $taskUtility(T_i, U_i)$. The agent will also know if there is any conflict $conflictingTasks(T_i, T_j)$ between tasks T_i, T_j. As different tasks are expected to be treated by different agents, the precedence $T_i(U_i) \prec T_k(U_k)$ informs an agent that the task T_k generates to the group a utility U_k greater than the utility U_i of the task T_i.

Primitive fluents are meant to define properties of the individual agents and the overall system that change in time.

Primitive Fluents

- $ownUtility(U_i)$ utility gained by the agent from achieving task T_i
- $instUtility(U_i)$ utility gained by the institution from achieving task T_i

Each agent has its $ownUtility(U_i)$, utility to measure the current benefit that the agent has gathered so far as reward to the work performed within the organisation. The gain obtained at the institution level is expressed by $instUtility(U_i)$. This is equivalent to the monetary gain that the institution can make in its own environment for the given work.

Defined fluents are very convenient to define properties which are not primitive, since they depend on other fluents.

Defined Fluents

- $prohibitingNorm(T_i, S) \equiv conflictingTasks(T_i, T_k) \wedge T_i(U_i) \prec T_k(U_k)$

When deciding whether to start acting for a given task, the agent should know whether there are conflicts or not.

Successor state axioms are used to specify the next state in the case that an action is executable.

Successor State Axiom

- $$\begin{aligned} Poss(a, S) \supset [state(I, do(a, S)) \equiv \\ &(a = actOnTask \wedge state(I, S)) \vee \\ &(a = senseTaskNorm \wedge state(Ip, S) \wedge I = Ip + 1) \vee \\ &(a = noAction \wedge state(I, S))] \end{aligned}$$

Here the state includes just the variable I, showing the timing when the agent has work to do, but more properties can be included, if relevant. With the current state $state(Ip, S)$ and action $senseTaskNorm$ the next state will be $state(I, do(a, S))$, where $I = Ip + 1$. Alternatively, with the same current state and $noAction$ to execute the next state will stay the same.

Finally, procedures represent a compact way to express more complex actions.

Golog Procedure

- Procedure specifying a cycle of generic agent:

```
proc execTask
    senseTaskNorm ;
    (actOnTask | noAction)
end
```

In this simple scenario the agent has to execute the procedure *execTask* at each moment of the environment time. It has to first *senseTask Norm,* that is to see if it has some work to do. Then the agent can choose between acting on the task or not. This choice is in fact decided on the basis of possibility of action execution specified by the precondition axioms defined in the next subsection.

3.5 Types of Institutionalized Agents

The precondition axioms define under what circumstances a given action can be executed. Here they are used to refine the generic agent into various kinds: social, rebellious, selfish.

Institutionalized Agents

- Social Agent:
  $Poss(actOnTask, S) \equiv reqTask(T_i) \wedge \neg\ prohibitingNorm(T_i, S)$
- Rebellious Agent:
  $Poss(actOnTask, S) \equiv reqTask(T_i)$
- Selfish Agent:
  $Poss(actOnTask, S) \equiv reqTask(T_i) \wedge taskUtility(T_i, U_i) \wedge U_i > 0$

The social agent of this scenario will execute his current task, if it exists, but only if there are no conflicts defined by the $prohibitingNorm(T_i, S)$ fluent. A rebellious agent will simply ignore any conflict that might occur and it will also disregard the utility gained. The selfish agent will do its task, if it exists, but only if it can benefit from it (it is interested just to increase its own utility). Note the transparency of defining these properties due to their specification in the logic of the fluent calculus.

4 Normative Positions in the Institution

There are many normative positions possible when dealing with a norm-based institution [21]. From the point of view of our scenario there are several participant agents, an administrator, and possibly the system, if we assume that malicious behaviour is sometimes possible. Assuming the nine participant agents illustrated in subsection 3.1 and using the permission operator *Perm,* one might encounter the normative positions illustrated below.

Normative Positions

- Participants:
  $\neg Perm(T_1) \wedge \neg Poss(T_1)$ for social agent P_1 at t=0
  $\neg Perm(T_9) \wedge Poss(T_1)$ for rebellious agent P_9 at t=0
  $\neg Perm(T_3) \wedge Poss(T_3)$ for selfish agent P_3 at t=1
- Administrator:
  $\neg Perm(T_1) \wedge Perm(T_2) \wedge Perm(T_4) \wedge Perm(T_5) \wedge \neg Perm(T_9)$ at t=0
  $Perm(T_1) \wedge \neg Perm(T_3) \wedge Perm(T_5) \wedge \neg Perm(T_8)$ at t=1
- System:
  $Perm(T_1) \wedge Perm(T_2) \wedge \neg Perm(T_3)$ at t=0
  $Perm(T_1) \wedge Perm(T_2) \wedge Perm(T_3) \wedge Perm(T_5) \wedge Perm(T_8)$ at t=1

Participant agents are interested in the permission for executing or not their own tasks. All three normative positions ilustrated above for the participant agents are permissible according to their character. The positions of the rebellious and selfish agents are not acceptable for the administrator. The administrator needs to consider more complex normative positions, like the ones shown at the times 0 and 1 of the environment time, which are both acceptable. Yet for the system things can get more complicated, the normative position shown for the time 0 being acceptable, while the one for time 1 suggesting intrusion of some kind. Its treatment might require the use of the deontic logic and the logic of action/agency for a set of regulations as defined in [21].

5 Discussion and Future Work

We have presented a specification for a model of a computational institution where agents can be endowed with some degree of freedom within norms imposed for the benefit of the overall institution. The model is implemented in Prolog [13] and future experiments are planned to show its flexibility for realistic implementations of multi-agent systems. In this model some artificial agents are delegated to act for human agents working within the organisation for which the multi-agent system is deployed. We have shown how an electronic institution [7,8,20,23] is populated with agents specified by an action theory [10,18], and how norms [21,22] can be used to improve the overall performance of the institution. We believe that such a framework is mandatory for tackling problems of social order in a computational society.

There are many approaches that are trying to model info-societies with social norms and reputation, including some that prefer emergence by some kind of evolution mechanism [3], as opposed to a designed institution. We have taken the last view since it seems a convenient path for current state of the art technology that might support near future deployment. Although the concept of e-institution is well known [6,7,8,20], we have found that the logic-based model defined in this paper offers a considerable advantage in its implementation and also in the formulation of norms to be included in the institution. Norm compliance has been studied in [15] by simulation experiments considering strategies used by agents in a norm-based system. While the specification of their agents is defined in Z, a software specification language, our logic-based specification in a Situation Calculus formalism is more direct and therefore the use of norms is more transparent.

The computational society framework presented in [1] uses the Event Calculus for reasoning about events with the aim to achieve an open computational system. Our model of e-institution allows also inter-agent communication, a very important feature when more elaborate forms of encounter between agents is envisaged. Alternatively, scenes can serve to represent various contexts that agents might prefer to use for interaction.

The next step in our future work will be to carry out experiments that show the social contribution and individual satisfaction for various sets of norms. As the e-institution model is based on communication we also intend to find what effects the norms have on the behavioural dynamics considering dialogue.

Various approaches are known that study the dynamics of institutionalized organisations, including the hierarchy among agents [9]. A relevant line of research, for such normative models, is to study the effects of the environmental factors on the agents' level of social consciousness [11].

Acknowledgements. We would like to thank the three anonymous referees for their helpful comments. Dr. Vasconcelos' work was sponsored by the European Union, contract IST-1999-10208, research grant **Sustainable Lifecycles of Information Ecosystems** (SLIE), while associated with the Division of

Informatics, University of Edinburgh, Edinburgh, UK. Dr. Letia would like to acknowledge the support of the European Commission through grant number HPRI-CT-1999-00026 (the TRACS Programme at EPCC, University of Edinburgh, UK), while part of his work, reported in this paper, has been done during a visit to the Department of Computer Science, National University of Singapore.

References

1. Artikis, A., Pitt, J., Sergot, M.: Animated specifications of computational societies. In Castelfranchi, C., Johnson, W.L., eds.: First International Joint Conference on Autonomous Agents and Multiagent Systems, Bologna, Italy (2002) 1053–1061
2. Carriero, N., Gelernter, D.: Linda in Context. Communications of the ACM **32** (1989) 444–458
3. Conte, R.: Emergent (info)institutions. Journal of Cognitive Systems Research **2** (2001) 97–110
4. Dignum, F., Morley, D., Sonenberg, E.A., Cavedon, L.: Towards socially sophisticated BDI agents. In: Proceedings of the 4th International Conference on Multi-Agent Systems, Boston, MA, IEEE Computer Society Press (2000) 111–118
5. Enderton, H.B.: A Mathematical Introduction to Logic. 2nd edn. Harcourt/Academic Press, Mass., USA (2001)
6. Esteva, M., de la Cruz, D., Sierra, C.: ISLANDER: an electronic institutions editor. In Castelfranchi, C., Johnson, W.L., eds.: First International Joint Conference on Autonomous Agents and Multiagent Systems, Bologna, Italy (2002) 1045–1052
7. Esteva, M., Padget, J., Sierra, C.: Formalizing a language for institutions and norms. In Meyer, J.J., Tambe, M., eds.: Intelligent Information Agents. LNAI 2333. Springer-Verlag (2002) 348–366
8. Esteva, M., Rodríguez-Aguilar, J.A., Sierra, C., Garcia, P., Arcos, J.L.: On the Formal Specification of Electronic Institutions. In Dignum, F., Sierra, C., eds.: Agent Mediated E-Commerce. Volume 1991 of LNAI. Springer-Verlag, Berlin (2001)
9. Gelati, J., Rotolo, A., Sartor, G., Governatori, G.: Actions, institutions, powers. Preliminary notes. In Paolucci, M., ed.: Regulated Agent-Based Social Systems: Theory and Applications. Springer-Verlag (2002) In this volume.
10. Giacomo, G.D., Lésperance, Y., Levesque, H.J.: ConGolog, a concurrent programming language based on the situation calculus. Artificial Intelligence **121** (2000) 109–169
11. Grosz, B.J., Kraus, S., Sullivan, D.G., Das, S.: The influence of social norms and social consciousness on intention reconciliation. Artificial Intelligence**142** (2002) 147–177
12. Halmos, P.R.: Naive Set Theory. Van Nostrand, Princeton, New Jersey (1960)
13. Intelligent Systems Laboratory: SICStus Prolog User's Manual. Swedish Institute of Computer Science, available at `http://www.sics.se/isl/` (2000)
14. Jonker, C.M., Letia, I.A., Treur, J.: Diagnosis of the dynamics within an organisation by trace checking of behavioural requirements. In Wooldridge, M., Ciancarini, P., Weiss, G., eds.: Agent-Oriented Software Engineering. LNCS 2222. Springer-Verlag (2002) 17–32
15. Lopez y Lopez, F., Luck, M., d'Inverno, M.: Constraining autonomy through norms. In Castelfranchi, C., Johnson, W.L., eds.: First International Joint Conference on Autonomous Agents and Multiagent Systems, Bologna, Italy (2002) 674–681

16. Manna, Z.: Mathematical Theory of Computation. McGraw-Hill Kogakusha, Ltd., Tokio, Japan (1974)
17. Pacheco, O., Carmo, J.: Role based model for the normative specification of organized collective agency and agents in interaction. Autonomous Agents and Multi-Agent Systems **6** (2003) 145–184
18. Reiter, R.: Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems. MIT Press (2001)
19. Reiter, R.: On knowledge-based programming in the situation calculus. ACM Transactions on Computational Logic **2** (2001) 433–457
20. Rodríguez-Aguilar, J. A. and Martín, F. J. and Noriega, P. and Garcia, P. and Sierra, C.: Towards a Formal Specification of Complex Social Structures in Multi-Agent Systems. In Padget, J., ed.: Collaboration between Human and Artificial Societies. Volume 1624 of LNAI. Springer-Verlag, Berlin (2000)
21. Sergot, M.: A computational theory of normative positions. ACM Transactions on Computational Logic **2** (2001) 581–622
22. Singh, M.P.: An ontology for commitments in multiagent systems: Towards a unification of normative concepts. Artificial Intelligence and Law **7** (1999) 97–113
23. Vasconcelos, W.W., Sabater, J., Sierra, C., Querol, J.: Skeleton-based agent development for electronic institutions. In Castelfranchi, C., Johnson, W.L., eds.: First International Joint Conference on Autonomous Agents and Multiagent Systems, Bologna, Italy (2002) 696–703
24. Vasconcelos, W.W., Robertson, D., Agusti, J., Sierra, C., Wooldridge, M., Parsons, S., Walton, C., Sabater, J.: A lifecycle for models of large multi-agent systems. In Ciancarini, P., Wooldridge, M., eds.: Agent-Oriented Software Engineering. LNCS 2222. Springer-Verlag (2002) 297–317
25. Vasconcelos, W.W.: Expressive global protocols via logic-based electronic institutions. In Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems, Melbourne, Australia (2003) to appear

A Model of Normative Multi-agent Systems and Dynamic Relationships

Fabiola López y López[1] and Michael Luck[2]

[1] Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla, México
fabiola@cs.buap.mx

[2] Department of Electronics and Computer Science
University of Southampton, UK
mml@ecs.soton.ac.uk

Abstract. For agents, one of the advantages of being part of a society is the satisfaction of those goals whose success depends on the abilities of other agents. In turn, societies are controlled by norms and, consequently, agents must be able first to model the society in which they exist, and then to identify the different relationships, due to norms, in which they might be involved in order to act appropriately. Both of these could mean the difference between the success or failure of their goals. To this end, this paper focuses on the identification of the basic components of norm-based systems, and on representing and analysing the *dynamic* relationships between member agents which result from the processing of norms.

1 Introduction

For agents, one of the advantages of being part of a society is the satisfaction of those goals whose success depends on the abilities of other agents. Since societies are controlled by norms, agents must be able to model their society, and they must be able to identify the different relationships, due to norms, in which they might be involved, in order to act appropriately. We argue that the correct identification of such relationships may be the difference between the success or failure of an agent's goals. For example, to select a plan, agents take into account not only their own obligations and prohibitions but also those of other agents, and to adopt a norm, agents must recognise an issuer's authority. Must research has been undertaken on how to incorporate norms into agents and multi-agent systems in taking steps toward the computational implementation of societies, institutions and organisations. This research has ranged from fundamental work on the importance of norms in agent behaviour [5,21], to proposing internal representations of norms [3,4,13], analysing the different types of norms [7,19], considering their emergence in groups of agents [22], proposing logics for their formalisation [18,23], and describing how agents manage norm adoption and compliance [2,8,14]. In the field of agents, research has primarily been focused at the level of multi-agent systems where norms represent the means to achieve

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 259–280, 2004.

coordination among their members. Current models of multi-agent systems regulated by norms assume not only that agents are able to comply with norms but also that they are able to obey the authorities of the system, mostly as an end in itself [6,9,11,16]. This means that authorities, norms and the relationships that arise from norms are all fixed at the start, so that the authority of agents can neither be objected to nor constrained (because the relationship is fixed). In this way, if an authority decides to apply punishments, an agent must accept those punishments even though it may consider that there is no motive to do so. This also constrains the flexibility of the system and loses the advantages that the autonomy of agents might provide. We argue that a model of multi-agent systems that considers the dynamism that arises from norms must be proposed. Moreover, since normative relationships that exist at one time may not last until another moment, agents must be provided with the means to identify changes in these relationships.

To date, existing models of system regulated by norms have not considered dynamic relationships between agents, yet these relationships can determine whether an agent will comply with norms. This paper address this limitation by proposing a model of multi-agent systems regulated by norms, and by describing the *dynamics* of norms and how, from the different stages in which norms are processed, different relations among agents can emerge. Besides the informal description, formal specifications of the main concepts and processes are given in order to avoid any ambiguity arising through the use of informal natural language. In particular, this avoids inconsistencies which might complicate the use and correct implementation of the theoretical framework provided.

In this document, first a general structure of a norm, and the basic characteristics of normative agents are discussed in Section 2. Then, in Section 3 a model of a multi-agent system regulated by norms is proposed. In the same section, some roles for agents that arise from norma, are identified. After that, the changes that occur in a system when norms are issued, complied with, or violated are described (Section 4). Finally, in Section 5, a set of normative relationships between agents is provided, before presenting our conclusions.

2 Norms and Normative Agents

In this section we describe the basic blocks from which to build up our theory of normative multi-agent systems. This conceptual infrastructure provides the basis for a broad theory, and underpins several aspects not included in this paper, but described elsewhere [14,15]. As a means to develop a formal model of a normative agent without repeating earlier work, we adopt the SMART agent framework described in [10]. In addition, in what follows, we also adopt the Z specification language to construct the formal model, because Z schemas allow, among other things, an easy transition from specifications to programs, there are tools that allow type checking, and so on. A Z schema contains two parts: the declaration part which declares local variables, and the predicate part which expresses some properties of the values of these variables. Z is based on set-theory and first order logic, with details available in [20]. For brevity, however, we will not elaborate the use of Z further.

2.1 Agents

In the SMART agent framework, an *attribute* represents a perceivable feature of the agent's environment which, here, is represented as a predicate or its negation. Then, a particular *state* in the environment is described by a set of attributes, a *goal* represents situations that an agent wishes to bring about, *motivations* are desires or preferences that affect the outcome of the reasoning intended to satisfy an agent's goals, and *actions* are discrete events that change the state of the environment when performed. For the purposes of this paper, we formally describe attributes, environmental states, goals and actions. Details of the remaining elements are not needed, so we simply consider them as given sets.

$$[Predicate, Motivation]$$

$$Attribute ::= pos\langle\langle Predicate\rangle\rangle \mid not\langle\langle Predicate\rangle\rangle$$

$$\begin{array}{ll} EnvState == \mathbb{P}_1\ Attribute & Goal == \mathbb{P}_1\ Attribute \\ Action == EnvState \nrightarrow EnvState & \end{array}$$

An autonomous agent is described by a non-empty set of attributes representing its permanent features, a set of goals that it wants to bring about, a set of capabilities that it is able to perform, and a non-empty set of motivations representing its preferences.

$$\begin{array}{|ll}
\multicolumn{2}{l}{AutonomousAgent} \\
\hline
capabilities : \mathbb{P}\ Action; & goals : \mathbb{P}\ Goal \\
motivations : \mathbb{P}\ Motivation; & beliefs : \mathbb{P}_1\ Attribute \\
\hline
goals \neq \varnothing; & motivations \neq \varnothing \\
\hline
\end{array}$$

2.2 Norms

An agent may have access to certain norms that are represented as data structures relating to social rules. These may be common to all agents (such as with a mutually understood social law) or only available to some. *Norms* are the mechanisms that a society uses in order to influence the behaviour of agents within it. Norms can be created from different sources, varying from built-in norms to simple agreements between agents, or more complex legal systems. They may persist during different periods of time; for example until an agent dies, as long as an agent remains in the society for which the norms were issued, or just for a short period of time until normative goals become satisfied. There are also different aspects that can be used to characterise them. First, norms are always prescribed to be complied with by a set of *addressee* agents in order to *benefit* another set of agents (possibly empty). They specify something that ought to be done, and consequently they include *normative goals* that must be satisfied by addressees. Sometimes, these normative goals must be directly intended, whereas at other times their role is to inhibit specific goals (as in the case of prohibitions). Second, norms are not always applicable, and their activation depends on the *context* in which agents are situated; there may be *exceptions* when agents are not obliged to comply with the norm. Finally, in some cases, norms suggest the existence of a set of *punishments* to be imposed when agents do not satisfy the normative goals, and a set of *rewards* to be received when agents do. Both punishment and rewards are also represented as sets of goals (which can be empty) that must be satisfied by someone else. Thus, the general structure of a norm can be formalised as follows. (Note that we specify normative goals as a set, to allow for the possibility of multiple goals in a norm, though we recognise that this will typically be a singleton set.)

$$\begin{array}{|l}
\multicolumn{1}{l}{\mathit{Norm}} \\
\hline
\mathit{addressees}, \mathit{beneficiaries} : \mathbb{P}\, \mathit{Agent} \\
\mathit{normativegoals}, \mathit{rewards}, \mathit{punishments} : \mathbb{P}\, \mathit{Goal} \\
\mathit{context}, \mathit{exceptions} : \mathit{EnvState} \\
\hline
\mathit{normativegoals} \neq \varnothing;\ \mathit{addressees} \neq \varnothing;\ \mathit{context} \neq \varnothing \\
\mathit{context} \cap \mathit{exceptions} = \varnothing;\ \mathit{rewards} \cap \mathit{punishments} = \varnothing \\
\hline
\end{array}$$

In the formalisation above, some constraints are imposed on the elements of a norm to eliminate the possibility of having norms that prescribe nothing, norms that are addressed to no one, norms that do not specify the situations in which they must be complied with, or norms that have inconsistencies in describing either the states in which agents are immune or between the rewards and punishments associated with a norm.

Norms can be divided, without eliminating the possibility of having further categories, into four types: *obligations, prohibitions, social commitments* and *social codes.* Broadly, we can say that *obligations* and *prohibitions* are norms adopted once an agent becomes a member of a society, *social commitments* are norms derived from agreements or negotiations between two or more agents, and *social codes* are norms motivated by feelings such as love, pity, friendship, or social conformity. It is not the purpose of this paper to discuss the different categories of norms but these can be found elsewhere [13]. In the remainder of this paper we will use the term *norm* as an umbrella term to cover every type of norm, even those that do not include punishments; although they can be created in different ways and with different purposes, we argue that all of them share the same structure. An important consideration at this point is that we understand *prohibitions* as norms whose normative goals must be avoided by addressee agents.

2.3 Normative Agents

We define a normative agent as an autonomous agent whose behaviour is shaped by the obligations it must comply with, the prohibitions that limit the kinds of goals that it can pursue, the social commitments that have been created during its social life, and social codes that may not carry punishments, but whose fulfillment is a means of being identified as part of a community.

$$\begin{array}{|l}
\hline
\mathit{NormativeAgent} \\
\hline
\mathit{AutonomousAgent} \\
\mathit{norms} : \mathbb{P}\, \mathit{Norm} \\
\hline
\end{array}$$

2.4 Permitted and Forbidden Actions

Sometimes, it is useful to observe norms not through the normative goals that ought to be achieved, but through the actions that can lead to the satisfaction of such goals. Then, we can consider actions that are either *permitted* or *forbidden* by a norm as follows. If there is a situation that makes a norm active, and the results of an action benefit the achievement of the associated normative goal, then such an action is *permitted* by the respective norm. For example, if the normative goal of a norm is to have taxes paid, then the action *paying taxes* is a permitted action if it changes an agent's situation of having taxes unpaid into a situation where taxes are paid. By analogy, we can define *forbidden* actions as those actions leading to a situation that contradicts or hinders the normative goal. For example, the action *illegal parking* is a forbidden action by a norm whose normative goal is to avoid parking in front of a hospital entrance. In general, it is not trivial to observe how the results of an action might benefit or hinder the achievement of normative goals. For instance, if we spend all our money and then try to pay our taxes, it might be not obvious that spending money may hinder the normative goal of paying taxes. To avoid drilling down into the intricate details of this important but somewhat secondary concern in relation to the focus of this paper, the associations between situation states that might *benefit* or *hinder* goals are taken for granted and formalised as follows.

$$\begin{array}{|l}
\mathit{benefited\_} : \mathbb{P}(\mathit{EnvState} \times \mathit{Goal}); \qquad \mathit{hindered\_} : \mathbb{P}(\mathit{EnvState} \times \mathit{Goal})
\end{array}$$

Now, we define two relations that hold among an action and a norm, which either permit or forbid the action, as follows.

$$\begin{array}{|l}
\mathit{permitted\_}, \mathit{forbidden\_} : \mathbb{P}(\mathit{Action} \times \mathit{Norm}) \\
\hline
\forall\, a : \mathit{Action};\ n : \mathit{Norm};\ \bullet \\
\mathit{permitted}\ (a, n) \Leftrightarrow (\exists\, g : n.\mathit{normativegoals} \bullet \mathit{benefited}\ (a\ n.\mathit{context}, g)) \land \\
\mathit{forbidden}\ (a, n) \Leftrightarrow (\exists\, g : n.\mathit{normativegoals} \bullet \mathit{hindered}\ (a\ n.\mathit{context}, g))
\end{array}$$

In other words, if an action is applied in the context of a norm, and the results of this action benefit the normative goals, then the action is permitted, otherwise the action is forbidden. In this way, norms act as action-filtering norms.

3 Normative Multi-agent Systems

Norms cannot be studied independently of the system for which they were created because they relate two or more members in a society, and it is the *social structure* that enforces norm compliance. Consequently, before describing how many processes due to norms are triggered, an analysis of the main components of a social system regulated by norms must be provided. A *normative multi-agent system* can be defined as a set of *normative* agents, which are controlled by a set of common *norms* ranging from obligations and social commitments to social codes. This *control* can be observed in three different aspects. First, member agents must recognise themselves as part of the society. Second, complete control cannot be exerted if sanctions or incentives are not applied when norms are either violated or complied with. Third, changes in the current normativity must be allowed as a way to solve unpredictable conflicts between agents and norms, or both. Each of these aspects is discussed in the subsections below.

3.1 Membership of Normative Societies

The performance of every structure of control relies on the capabilities of its members to recognise and follow its norms. However, since agents are autonomous, the fulfillment of norms can never be taken for granted. In fact, autonomous agents decide whether to comply with norms based on their own current goals and motivations [14]. It is also possible that not all the norms that one agent has adopted belong to just one system, because agents can be part of more than one society at the same time. In addition, due to agent limitations, it is possible that not all the norms of the system are known by any agent. These characteristics can be formally expressed by saying that the set of norms adopted by any member is not necessarily a subset of the norms of the system, and also that the intersection of both sets of norms is not empty. Now, part of being a member of a society means that agents are subject to some of the norms in the system. In other words, the set of addressee agents of every norm must be included in the set of member agents, because it does not make any sense to have norms addressed to non-existent agents.

3.2 Interlocking Norms

The norms of a system are not isolated from each other; sometimes, compliance with them is a condition to trigger (or activate) other norms. That is, there are norms that prescribe how some agents must behave in situations in which other agents either comply with a norm or do not comply with it [17]. For example, when employees comply with their obligations in an office, paying their salary becomes an obligation of the employer; or when a plane cannot take-off, providing accommodation to passengers becomes a responsibility of the airline. Norms related in this way can make a complete chain of norms, because the newly activated norms can, in turn, activate new ones. Now, since triggering a norm depends on past compliance with another norm, we call these kinds of norms *interlocking norms.* The norm that gives rise to another norm is called the *primary* norm, whereas the norm activated as a result of either the fulfillment or violation of the first is called the *secondary* norm. In terms of the norm model mentioned earlier, the *context* is a state that must hold for a norm to be complied with. Since the fulfillment of a norm is assessed through its normative goals, the context of the secondary norm must include the satisfaction (or non-satisfaction) of all the primary norm's normative goals. Figure 1 illustrates the structure of both the primary and the secondary norms and how they are interlocked through the primary norm's normative goals and the secondary norm's context. To formalise this kind of norm, some definitions are needed. We say that a norm can be considered as *fulfilled* in a specific state of the system if its corresponding normative goals are a logical consequence of such a state.

$$\begin{array}{|l}
\mathit{logicalconsequence}\_ : \mathbb{P}(\mathit{EnvState} \times \mathit{EnvState}) \\
\mathit{fulfilled}\_ : \mathbb{P}(\mathit{Norm} \times \mathit{EnvState}) \\
\hline
\forall\, n : \mathit{Norm};\ st : \mathit{EnvState} \bullet \\
\quad \mathit{fulfilled}\ (n, st) \Leftrightarrow (\forall\, g : n.\mathit{normativegoals} \bullet \mathit{logicalconsequence}\ (st, g))
\end{array}$$

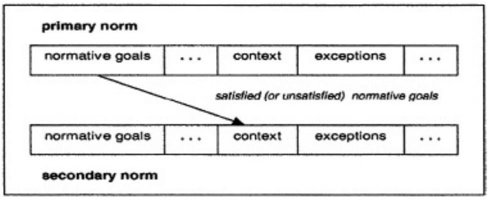


Fig. 1. Interlocking Norm Structure

Formally, a norm is interlocked with another norm *by non-compliance* if, in the context of the secondary norm, an instance of the primary norm can be considered as violated. This means that when any addressee of a norm does not fulfill the norm, the corresponding interlocking norm will be triggered. The formal specification of this is given below. There, n_1 represents the primary norm, whereas n_2 is the secondary norm.

$$\begin{array}{|l}
\mathit{lockedbynoncompliance}\_ : \mathbb{P}(\mathit{Norm} \times \mathit{Norm}) \\
\hline
\forall\, n_1, n_2 : \mathit{Norm} \bullet \quad \mathit{lockedbynoncompliance}\ (n_1, n_2) \Leftrightarrow \\
\quad (\exists\, n_1 : \mathit{Norm} \mid\bullet \neg\, \mathit{fulfilled}\ (n_1, n_2.\mathit{context}))
\end{array}$$

Similarly, a norm is interlocked with another norm *by compliance* if, in the context of the secondary norm, an instance of the primary norm can be considered as fulfilled. Thus, any addressee of the norm that fulfills it will trigger the interlocking norm. The specification of this is given as follows.

$$\begin{array}{|l}
\mathit{lockedbycompliance\_} : \mathbb{P}(\mathit{Norm} \times \mathit{Norm}) \\
\hline
\forall n_1, n_2 : \mathit{Norm} \bullet \quad \mathit{lockedbycompliance}\ (n_1, n_2) \Leftrightarrow \\
\qquad (\exists n_2 : \mathit{Norm} \mid \mathit{fulfilled}\ (n_1, n_2.\mathit{context}))
\end{array}$$

3.3 Enforcement and Encouragement of Norm Compliance

Complete control cannot be exerted if, for each norm in the system, there is no other norm that prescribes how some agents have to react when the original norm is violated [17]. For example, if there is an obligation to pay accommodation fees for all students in a university, there must also be a norm stating what hall managers must do when a student does not pay. These norms are addressed to a specific group of agents responsible for punishing non-compliance. It is only through these norms that some agents are entitled to punish other agents. Chaos might emerge in a society if such responsibility is given either to no one or to anyone. Addressee agents of this kind of norm are frequently called the *defenders* of a norm.

To describe these norms, we observe that the violation of a norm can be detected by an agent when it realises that the associated normative goals were not satisfied. Once this event becomes identified by defenders, their duty is then to start a process in which rebellious agents can be punished. This suggests that these norms can be modelled as interlocking norms with the additional restriction that every punishment included in the violated norm must appear in the normative goals of the secondary norm. That is, defenders of norms must have the goal of punishing every offender of a norm. Figure 2 shows how both the structure of a norm and the norm which enforces it, are related. Formally, a relationship between a norm directed to control the behaviour of agents and a secondary norm can be defined as follows. A norm *enforces* another norm through punishments if they are interlocked by non-compliance, and the punishments associated with the unfulfilled norm are part of the normative goals of the enforcement norm. We call these kinds of norms *enforcement* norms.

$$\begin{array}{|l}
\mathit{enforce\_} : \mathbb{P}(\mathit{Norm} \times \mathit{Norm}) \\
\hline
\forall n_1, n_2 : \mathit{Norm} \bullet \mathit{enforce}\ (n_1, n_2) \Leftrightarrow \\
\qquad \mathit{lockedbynoncompliance}(n_1, n_2) \wedge n_2.\mathit{punishments} \subseteq n_1.\mathit{normativegoals})
\end{array}$$

So far, we have described secondary norms in term of punishments because punishments are one of the more commonly used mechanisms to enforce compliance with norms. However, a similar analysis can be undertaken for secondary norms corresponding to the process of rewarding members fulfilling their duties. The relations between norms and norms to reward their compliance are shown in Figure 3. Formally, we say that a norm *encourages* compliance with another norm through rewards if they are locked by compliance, and the rewards associated with the fulfilled norm are part of the normative goals of the *encourage* norm. We call these kinds of norms *reward* norms.

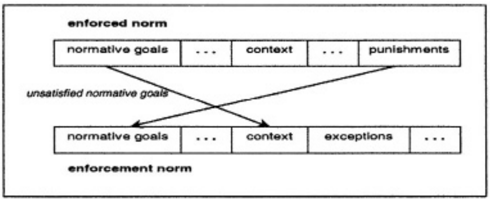


Fig. 2. Enforcement Norm Structure

$$reward\_ : \mathbb{P}(Norm \times Norm)$$

$$\forall n_1, n_2 : Norm \bullet reward\ (n_1, n_2) \Leftrightarrow$$
$$lockedbynoncompliance(n_1, n_2) \wedge n_2.rewards \subseteq n_1.normativegoals)$$

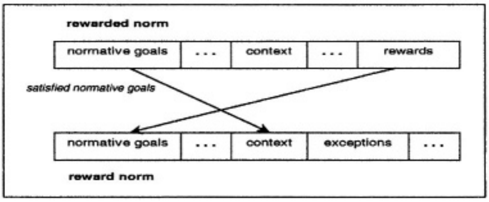


Fig. 3. Reward Norm Structure

Now, it is important to state that this way of representing enforcement norms can create an infinite chain of norms because we would also have to define norms to use when authorities or defenders do not comply with their obligations to either punish those agents breaking rules or reward those agents who fulfill their responsibilities [17]. To avoid this chain of norms, and by taking the risk of being considered as absolutist, in what follows we consider that no punishments are applied when an enforcement norm is not fulfilled. This means that neither authorities nor defenders can be judged (at least in this normative system) by dismissing their responsibilities. (A similar analysis can be undertaken for reward norms.) Nevertheless, if required, our model and formalisation for enforcing and encouraging norms can be used recursively as necessary. There is nothing in the definition of the model itself to prevent this.

3.4 Dynamic Normativity and Legislation

In general, norms are introduced into a society as a means to achieve social order. Some norms are intended to avoid conflicts between agents, others to allow the establishment of commitments, and others still to unify the behaviour of agents as a means of social identification. However, neither all conflicts nor

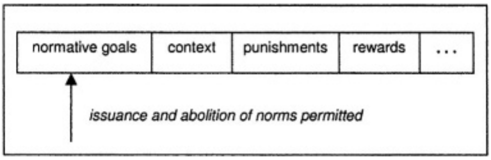


Fig. 4. Legislation Norm Structure

all commitments can be anticipated and controlled by norms. Consequently, in a *dynamic* multi-agent system there must exist the possibility of creating new norms, modifying existing norms, or even abolishing those that become obsolete. Now, although it is possible that many of the members of a society have capabilities to do this, these actions must be restricted to be carried out by a particular set of agents in a particular situation in order to avoid anyone imposing its norms, because some conflicts of interest might emerge. In other words, norms stating when actions to legislate are permitted must be also included [12] . These norms are called *legislation* norms, and they must specify that actions to issue and abolish norms are only permitted by a particular set of agents represented in its addressees (see Fig. 4). These constraints are specified below.

$$\begin{array}{|l}
legislate\_ : \mathbb{P}\, Norm \\
\hline
\forall\, n : Norm \bullet legislate\ (n) \Leftrightarrow (\exists\, issuingnorms, abolishnorms : Action \bullet \\
\quad permitted\ (issuingnorms, n) \vee permitted\ (abolishnorms, n))
\end{array}$$

3.5 Formal Model

All the elements discussed above are now incorporated into the formal representation of a *normative multi-agent system,* given below.

$$\begin{array}{|l}
\hline
NormativeMAS \\
\hline
members : \mathbb{P}\, NormativeAgent; \qquad normsNMAS : \mathbb{P}\, Norm \\
enforcenorms : \mathbb{P}\, Norm; \qquad rewardnorms : \mathbb{P}\, Norm \\
legislationnorms : \mathbb{P}\, Norm \\
\hline
\forall\, ag : members \bullet ag.norms \cap normsNMAS \neq \varnothing \\
\forall\, rg : normsNMAS \bullet rg.addressees \subseteq members \\
\forall\, en : enforcenorms \bullet (\exists\, n : normsNMAS \bullet enforce\ (en, n)) \\
\forall\, rn : rewardnorms \bullet (\exists\, n : normsNMAS \bullet reward\ (rn, n)) \\
\forall\, ln : legislationnorms \bullet legislate\ (ln) \\
\hline
\end{array}$$

That is, a normative multi-agent system comprises the following elements: a set of member agents able to reason about norms, a set of norms directed to regulate the behaviour of these agents (represented here by the variable *normsNMAS*), a set of norms directed to enforce and judge the latter set of norms (*enforcenorms*), the set of norms directed to encourage compliance with norms through rewards (*rewardnorms*), and the norms issued to allow the creation and abolition of norms (*legislationnorms*). In the schema, the first predicate states that all members must have adopted some of the norms of the normative multi-agent system, and the second makes explicit that addressee agents of this set of norms must be members of the system. The last three predicates respectively describe the structure of enforcement, reward and legislation norms. Notice that whereas every enforcement norm must have a norm to enforce, not every norm may have a corresponding enforcement norm, which means that no one in that society is legally entitled to punish an agent that does not fulfill such a norm.

3.6 Normative Roles

Defining a normative multi-agent system in this way allows the identification of different roles for agents that depend on the kinds of norms agents are responsible for. Specifically, are agents entitled to create, modify, or abolish the set of norms of a society. No other members of the society are endowed with the power and authority to do so. These kinds of agents can, in turn, be either elected or decreed, and we call them *legislators*. An agent is an *defender* if it is directly responsible for either applying punishments or giving rewards. *Addressee* agents are directly responsible for the achievement of normative goals, and *beneficiaries* are agents whose goals can benefit when a normative goal becomes satisfied. Both addressees and beneficiaries can be directly identified from the model of norms. To identify defenders and legislators we need the following relations. The first and second relations state which agents are entitled to punish or reward a norm in a normative multi-agent system. The third relation specifies which agents can be considered the defenders of a particular norm. Finally, the fourth relation states who is a legislator.

$$
\begin{array}{|l}
\mathit{canpunish}\_ : \mathbb{P}(\mathit{NormativeAgent} \times \mathit{Norm} \times \mathit{NormativeMAS}) \\
\mathit{canreward}\_ : \mathbb{P}(\mathit{NormativeAgent} \times \mathit{Norm} \times \mathit{NormativeMAS}) \\
\mathit{isdefender}\_ : \mathbb{P}(\mathit{NormativeAgent} \times \mathit{Norm} \times \mathit{NormativeMAS}) \\
\mathit{islegislator}\_ : \mathbb{P}(\mathit{NormativeAgent} \times \mathit{NormativeMAS}) \\
\hline
\quad \forall\, ag : \mathit{NormativeAgent};\ n : \mathit{Norm};\ nmas : \mathit{NormativeMAS} \bullet \\
\qquad \mathit{canpunish}\ (ag, n, nmas) \Leftrightarrow (n \in nmas.\mathit{normsNMAS} \wedge \\
\qquad\quad (\exists\, en : \mathit{Norm} \bullet (en \in nmas.\mathit{enforcenorms} \wedge \\
\qquad\qquad\quad ag \in en.\mathit{addressees} \wedge \mathit{enforce}\ (en, n)))) \wedge \\
\qquad \mathit{canreward}\ (ag, n, nmas) \Leftrightarrow (n \in nmas.\mathit{normsNMAS} \wedge \\
\qquad\quad (\exists\, en : \mathit{Norm} \bullet (en \in nmas.\mathit{enforcenorms} \wedge \\
\qquad\qquad\quad ag \in en.\mathit{addressees} \wedge \mathit{reward}\ (en, n)))) \wedge \\
\qquad \mathit{isdefender}\ (ag, n, nmas) \Leftrightarrow (\mathit{canpunish}\ (ag, n, nmas) \vee \\
\qquad\qquad \mathit{canreward}\ (ag, n, nmas)) \wedge \\
\qquad \mathit{islegislator}\ (ag, nmas) \Leftrightarrow (\exists\, ln : \mathit{Norm} \bullet \\
\qquad\qquad ln \in nmas.\mathit{legislationnorms} \wedge\ ag \in ln.\mathit{addressees})
\end{array}
$$

These *normative roles* for agents are not mutually exclusive. Agents are able to have more than one normative role at the same time, depending on the kind of norm being considered. For example, in a social commitment, the beneficiary agent can be a defender and consequently encourage the fulfillment of a norm, and either apply punishments or give rewards. In an office, the manager can be both a legislator and then impose his own norms, and a defender entitled to punish his employees. The more complex a society, the more elaborate these normative roles become, and in some cases all legislators, authorities (judges), and police make a complex structure of control generally named *government,* with its own legal norms directed at controlling the rest of the society. Thus, being a defender is a relationship that holds between an agent and the enforcement norm that entitles it to defend the norm. Similarly, being a legislator means that there is a norm that entitles an agent to modify the current legislation by creating new norms and abolishing some of the norms already created. Considering defenders and legislators in this way allows us to represent the fact that all these elements cannot be taken independently of each other, but are complementary.

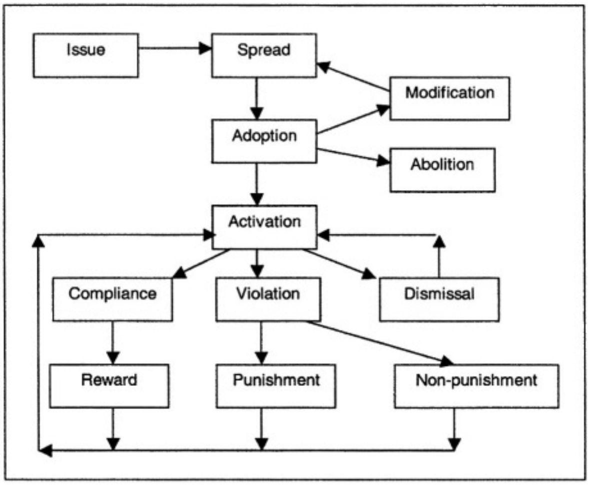


Fig. 5. Norm Dynamics

4 Dynamics of Norms

Norms are not a static concept. Once they are included in a system, they cause certain behaviour in each of the members. In Figure 5, the different processes through which a norm passes from its creation to its abolition can be observed. Arrows represent the transitions from one stage of norms to another. That is, first a legislator issues a norm. After that, the norm is spread by either indirect or direct communication. Then, adoption of norms takes place. Through this process, an agent expresses its willingness to comply with the norm as a way of being part of the society. Once a norm is adopted, it remains inactive, or in *latency,* until the applicability conditions are satisfied. In exception states, agents are not obliged to comply with norms, and consequently norms can be dismissed. However, in the majority of the cases, two different situations might occur after a norm becomes activated: a norm is either fulfilled or violated by addressee agents. After a norm is complied with, a reward can be offered. By contrast, if the norm is violated there are two possibilities: either punishments are applied or they are not. Finally, as time progresses, some norms become abolished or modified.

Considering the *dynamics* that result from norms is an important issue that deserves our attention, because interesting relations among agents can be identified in each (see Section 5). In turn, according to these relationships, different reactions of agents are expected. For instance, when a norm is activated, defenders are just entitled to require its fulfillment. However, if the norm is violated, defenders are entitled to apply punishments. Consequently, we argue that the correct identification of the different stages of a norm is key to modelling the normative behaviour of agents. In the following subsections, the transitions between these different stages are described and formalised from the perspective of an external observer.

4.1 Changing Legislation

Legislation of norms is a responsibility only attributed to legislator agents. Such a responsibility comprises at least three functions, namely issuance, abolition, and modification of norms. Unfortunately, due to their complexity, details of how such functions are carried out cannot be given here. Determining how, when and why a norm must be created requires a complex analysis of the current conditions of a system. However, we can still introduce two functions to identify the recently created norms (*newnorms*), and the norms that must be abolished (*obsoletenorms*). These functions might be equivalent to asking legislators about the results of their assigned tasks. In turn, modification of norms can be seen as the abolition of a subset of norms together with the issuance of another subset of norms with the same name, so a specific function to do that is not included here. Now, after new norms are created and others are abolished, the spreading and updating of norms is needed. As a result of these changes at global level, the set of member agents must also change. That is, some of these norms become internally adopted or abolished by addressee agents. This is represented by the functions *spreadnorms* and *abolishnorms*, which can be seen as the processes through which agents are notified of the creation of new norms and the abolition of norms that become obsolete. The *NormLegislation* schema formalises the functions associated with the legislation of norms, in which the variable *nmas* represents the normative multi-agent system in which changes in legislation occur.

$$\begin{array}{|l}
\hline \mathit{NormLegislation} \\
\mathit{nmas} : \mathit{NormativeMAS}; \qquad \mathit{legislators} : \mathbb{P}\, \mathit{NormativeAgent} \\
\mathit{newnorms}, \mathit{obsoletenorms} : \mathbb{P}\, \mathit{NormativeAgent} \rightarrow \mathbb{P}\, \mathit{Norm} \\
\mathit{spreadnorms} : (\mathbb{P}\, \mathit{NormativeAgent} \times \mathbb{P}\, \mathit{Norm}) \rightarrow \mathbb{P}\, \mathit{NormativeAgent} \\
\mathit{abolishnorms} : (\mathbb{P}\, \mathit{NormativeAgent} \times \mathbb{P}\, \mathit{Norm}) \rightarrow \mathbb{P}\, \mathit{NormativeAgent} \\
\hline \quad \forall\, \mathit{ag} : \mathit{legislators} \bullet \mathit{islegislator}\ (\mathit{ag}, \mathit{nmas}) \\
\quad \mathrm{dom}\ \mathit{newnorms} = \mathbb{P}\, \mathit{legislators}; \qquad \mathrm{dom}\ \mathit{obsoletenorms} = \mathbb{P}\, \mathit{legislators} \\
\hline
\end{array}$$

Now, the process that changes norms in both the system and its members can be represented as follows.

$$\begin{array}{|l}
\hline \mathit{ChangeLegislation} \\
\Delta \mathit{NormLegislation} \\
\hline \quad \mathit{nmas}'.\mathit{normsNMAS} = \mathit{nmas}.\mathit{normsNMAS} \setminus \\
\qquad\qquad \mathit{obsoletenorms}\ \mathit{legislators} \cup \mathit{newnorms}\ \mathit{legislators} \\
\quad \mathit{nmas}'.\mathit{members} = \mathit{spreadnorms}\ (\mathit{abolishnorms}\ (\mathit{nmas}.\mathit{members}, \\
\qquad\qquad \mathit{obsoletenorms}\ \mathit{legislators}), \mathit{newnorms}\ \mathit{legislators}) \\
\hline
\end{array}$$

The first predicate states that the set of norms, after a change in legislation, is composed of all the old norms except those recently abolished, combined with the recently created norms. The second predicate represents how all members are informed of legislation changes through a composition of functions. That is, first members are informed about norms that must be abolished because they are now considered obsolete, and then they receive information about the recently created norms.

4.2 Normative Multi-agent System State

After norms are issued, spread, and then adopted, they enter in a cycle in which different agents intervene. To capture the different stages in which a norm is processed, we specify the *state* of a normative multi-agent system as follows.

$$\begin{array}{|l}
\hline \mathit{NMASState} \\
\mathit{NormativeMAS} \\
\mathit{currentsituation} : \mathit{EnvState} \\
\mathit{formeractivenorms}, \mathit{activenorms}, \mathit{fulfillednorms} : \mathbb{P}\, \mathit{Norm} \\
\mathit{unfulfillednorms}, \mathit{punishednorms}, \mathit{rewardednorms} : \mathbb{P}\, \mathit{Norm} \\
\hline \quad \mathit{activenorms} \subseteq \mathit{normsNMAS} \\
\hline
\end{array}$$

At a particular instant of time, some norms become *activated.* This means that the conditions under which a norm must be fulfilled are satisfied. Moreover, other previously activated norms become either *fulfilled* or *unfulfilled.* Furthermore, some of the unfulfilled norms become *punished,* and some of the fulfilled ones become *rewarded.* Identifying these stages of norms is important because any change in them can cause reactions in other agents. For example, addressee agents acquire new responsibilities because of active norms, and they deserve to be rewarded or punished due to fulfilled or unfulfilled norms respectively. In addition, some agents might require compliance with active norms, or apply punishments to addressees of unfulfilled norms, etc. In the schema for the state of a normative multi-agent system, the *formeractivenorms* variable represents the norms that were activated previously, and the *currentsituation* represents the state of the general environment.

4.3 Assessing Compliance with Norms

Although not all norms change their stage at the same time, we take a particular point in the time to assess them all. Now, as mentioned before, the easy way to determine if a norm has been fulfilled is by observing the current state of the system and then verifying if the associated normative goals are satisfied. This form of verifying compliance with norms can be used for any kind of norm, ranging from the norms of the normative system to the norms to enforce compliance with. These changes are represented in the schema below.

$$
\begin{array}{|l}
\textit{AssessNorm} \\
\hline
\textit{NormativeMAS}; \qquad \Delta \textit{NMASState} \\
\textit{observedchanges} : \textit{EnvState} \nrightarrow \textit{EnvState} \\
\textit{newactive}, \textit{newfulfilled} : \mathbb{P}\, \textit{Norm} \\
\textit{newpunished}, \textit{newrewarded} : \mathbb{P}\, \textit{Norm} \\
\hline
\quad \textit{currentsituation}' = \textit{observedchanges} \;\; \textit{currentsituation} \\
\quad \textbf{let } \textit{newactive} == \{ n : \textit{normsNMAS} \mid \\
\qquad \textit{logicalconsequence}\ (\textit{currentsituation}', n.\textit{context}) \} \bullet \\
\quad \textbf{let } \textit{newfulfilled} == \{ n : \textit{activenorms} \mid \textit{fulfilled}\ (n, \textit{currentsituation}') \} \bullet \\
\quad \textbf{let } \textit{newpunished} == \{ n : \textit{unfulfillednorms} \mid (\exists\, \textit{en} : \textit{enforcenorms} \bullet \\
\qquad (\textit{enforcepunish}\ (\textit{en}, n) \land \textit{fulfilled}\ (\textit{en}, \textit{currentsituation}'))) \} \bullet \\
\quad \textbf{let } \textit{newrewarded} == \{ n : \textit{fulfillednorms} \mid (\exists\, \textit{en} : \textit{enforcenorms} \bullet \\
\qquad (\textit{enforcereward}\ (\textit{en}, n) \land \textit{fulfilled}\ (\textit{en}, \textit{currentsituation}'))) \} \bullet (\\
\qquad\quad \textit{formeractivenorms}' = \textit{formeractivenorms} \cup \\
\qquad\qquad\quad \textit{activenorms} \setminus \textit{newactive} \land \\
\qquad\quad \textit{activenorms}' = \textit{newactive} \land \\
\qquad\quad \textit{fulfillednorms}' = \textit{fulfillednorms} \cup \textit{newfulfilled} \land \\
\qquad\quad \textit{unfulfillednorms}' = \textit{unfulfillednorms} \cup \\
\qquad\qquad\quad (\textit{activenorms} \setminus \textit{newfulfilled}) \land \\
\qquad\quad \textit{punishednorms}' = \textit{punishednorms} \cup \textit{newpunished} \land \\
\qquad\quad \textit{rewardednorms}' = \textit{rewardednorms} \cup \textit{newrewarded}) \\
\hline
\end{array}
$$

In this schema, *observedchanges* is a function that reports the observed changes in the social environment, and can be used to update the sets of norms. First, the set of new *active* norms is calculated by analysing if the context to trigger a norm, is a *logical consequence* of the current situation of the system. After that, the set of active norms that were *fulfilled* by their corresponding addressee agents is calculated by verifying the satisfaction of the corresponding normative goals. Next, unfulfilled norms that were *punished* are found by verifying if the norm that enforces it has already been satisfied. Verifying if fulfilled norms were *rewarded* is done similarly. After all these steps, the states of norms are updated accordingly.

5 Normative Relationships

As stated earlier, norms at different states create different kinds of relationships among agents. We identify four sets. The first is created due to the authority of certain agents in the system. The next is created once norms become activated. Norms that have been complied with also generate relations among agents through offered rewards. Finally, violated norms, and their associated punishments, cause agents to be related in a different way. These relationships are used by agents when reasoning about norms is needed, and a decision must be taken. Then, by using the proposed structure of the norm, the definition of a normative multi-agent system, and the different normative roles that agents might have be performing, we describe the set of relationships we are interested in. These relations are illustrated in Figure 6 in which rounded boxes represent the state of a norm, and hexagons symbolize the relationships created by them.

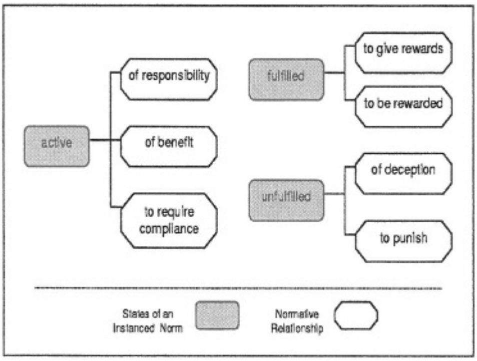


Fig. 6. Normative Relationships

5.1 Legislation Relations

As stated earlier, not all agents in a normative multi-agent system are entitled to legislate and, therefore, before a norm is adopted, agents must recognise the authority of the issuer, otherwise the validity of the norm could be questioned, and then rejected. Formally, we say that an agent is a *legal authority* for another agent if it is a legislator in the normative multi-agent system to which the second agent belongs.

$$\begin{array}{|l}
legalauthority\_ : \mathbb{P}(NormativeAgent \times NormativeAgent \times NormativeMAS) \\
\hline
\forall ag_1, ag_2 : NormativeAgent;\ nmas : NormativeMAS \bullet \\
\quad legalauthority(ag_1, ag_2, nmas) \Leftrightarrow \\
\qquad islegislator(ag_1, nmas) \wedge ag_2 \in nmas.members
\end{array}$$

5.2 Active Norm Relations

Norms become *activated* when the current situations of an agent (or a group of agents) match the context in which a norm must be fulfilled. For example, if a driver wants to park its car in front of an entrance, the norm that forbids such an action is applied, otherwise agents do not need to be concerned with them. ¿From this situation, four relations among agents can be inferred as follows.

It can be observed that some norms include exception states in which an agent is not obliged to respect those norm. An exception state could be treated as a state not included in the context of a norm, because in that case the norm would not be activated and, consequently, agents would not be obliged to comply. Although the results are similar, we prefer to make them explicit because it allows an agent to explain why it is not obliged to comply with that norm. This latter aspect can be useful if the norm is addressed to a set of agents, only some of which are excepted from their responsibilities. Formally, we say that an agent *can dismiss* a norm in a particular state of the system if that agent is an addressee of the norm, and the exception states of the norm are a logical consequence of the current state.

$$\begin{array}{|l}
candismiss\_ : \mathbb{P}(NormativeAgent \times Norm \times EnvState) \\
\hline
\forall ag_1, ag_2 : NormativeAgent;\ n : Norm;\ nmas : NormativeMAS; \\
\quad st : EnvState \bullet candismiss\ (ag_1, n, st) \Leftrightarrow \\
\qquad (\ ag_1 \in n.addressees \wedge logicalconsequence\ (st, n.exceptions))
\end{array}$$

Another important relationship that can be observed here is the relation between an addressee agent, a norm, and its defender. In this situation, it can be said that an agent is entitled to require compliance with norms either by threatening agents with future punishments, or by offering future rewards. Formally, it can be said that an agent *can require* another agent to fulfill a norm if it is a designated defender in the system, and the second agent is an addressee of the norm.

$$\begin{array}{|l}
canrequire\_ : \mathbb{P}(NormativeAgent \times NormativeAgent \times Norm \\
\qquad\qquad \times NormativeMAS) \\
\hline
\forall ag_1, ag_2 : NormativeAgent;\ n : Norm;\ nmas : NormativeMAS; \\
st : EnvState \bullet \\
\quad canrequire\ (ag_1, ag_2, n, nmas) \Leftrightarrow (isdefender\ (ag_1, n, nmas) \wedge \\
\qquad ag_2 \in nmas.members \wedge\ ag_2 \in n.addressees)
\end{array}$$

Finally, there are two further important relationships between agents. The first is the responsibility that an addressee agent has as soon as a norm becomes activated. Note that although an agent has a responsibility to fulfill, it does not means that it is going to do so. The decision is only made by the agent itself. Formally, we say that an agent *has a responsibility* to another if there is a norm already addressed to the first agent, and the benefits may be enjoyed by the second.

$$
\begin{array}{|l}
hasresponsibility\_ : \mathbb{P}(NormativeAgent \times NormativeAgent \times Norm \\
\qquad\qquad\qquad \times NormativeMAS) \\
\hline
\forall\, ag_1, ag_2 : NormativeAgent;\ n : Norm;\ nmas : NormativeMAS; \\
st : EnvState \bullet \\
\quad hasresponsibility\ (ag_1, ag_2, n, nmas) \Leftrightarrow (n \in nmas.normsNMAS \land \\
\qquad ag_1 \in nmas.members \land\ ag_2 \in nmas.members \land \\
\qquad ag_1 \in n.addressees \land\ ag_2 \in n.beneficiaries)
\end{array}
$$

The second relationship is its counterpart which relates to the expectations of a beneficiary agent to receive something from the responsibilities of others. Formally we say that an agent *expects benefits* from the responsibility of another agent if the former is the beneficiary of a norm addressed to the second agent.

$$
\begin{array}{|l}
expectsbenefit\_ : \mathbb{P}(NormativeAgent \times NormativeAgent \times Norm \times \\
\qquad\qquad\qquad NormativeMAS) \\
\hline
\forall\, ag_1, ag_2 : NormativeAgent;\ n : Norm;\ nmas : NormativeMAS; \\
st : EnvState \bullet \\
\quad expectsbenefit\ (ag_1, ag_2, n, nmas) \Leftrightarrow (n \in nmas.normsNMAS \land \\
\qquad ag_1 \in nmas.members \land ag_2 \in nmas.members \land \\
\qquad ag_2 \in n.addressees \land\ ag_1 \in n.beneficiaries)
\end{array}
$$

5.3 Fulfilled Norm Relations

Once a norm is *fulfilled,* no further action is necessary except maybe by addressee agents claiming rewards from a defender. Then, two complementary relationships are identified as follows. First, we say that an agent has the responsibility of *rewarding* another agent if the first agent is a defender of the norm and the second is an agent who has fulfilled it. In addition, an agent has the right to be *rewarded* by a defender of a norm if the first agent has already complied with it.

$$\begin{array}{l}
\mathit{rewards\_} : \mathbb{P}(\mathit{NormativeAgent} \times \mathit{NormativeAgent} \times \mathit{Norm} \times \mathit{EnvState} \\
\qquad\qquad \times \mathit{NormativeMAS}) \\
\mathit{rewarded\_} : \mathbb{P}(\mathit{NormativeAgent} \times \mathit{NormativeAgent} \times \mathit{Norm} \times \mathit{EnvState} \\
\qquad\qquad \times \mathit{NormativeMAS}) \\
\hline
\forall\, ag_1, ag_2 : \mathit{NormativeAgent};\ n : \mathit{Norm};\ nmas : \mathit{NormativeMAS}; \\
st : \mathit{EnvState} \bullet \\
\quad \mathit{rewards}\ (ag_1, ag_2, n, st, nmas) \Leftrightarrow (\ ag_2 \in n.\mathit{addressees} \land \\
\qquad ag_2 \in nmas.\mathit{members} \land \mathit{canreward}\ (ag_1, n, nmas) \land \mathit{fulfilled}\ (n, st)) \land \\
\quad \mathit{rewarded}\ (ag_1, ag_2, n, st, nmas) \Leftrightarrow (\ ag_1 \in n.\mathit{addressees} \land \\
\qquad ag_1 \in nmas.\mathit{members} \land \mathit{canreward}\ (ag_2, n, nmas) \land \mathit{fulfilled}\ (n, st))
\end{array}$$

5.4 Unfulfilled Norm Relations

By contrast, when a norm is *vilated,* several events take place and other kinds of relationships hold. Obviously, addressees of an unfulfilled norm will do nothing, and would prefer that their failure remains hidden in order to avoid facing the consequences of their actions. However, a *deception* situation emerges in which the interests of third agents (the beneficiaries) might be badly affected by the irresponsibility of offenders. Agents in this situation could claim compensation. Formally, it can be said that an agent is *deceived by* another agent if a norm was violated by the second agent, and the benefits could have been enjoyed by the first.

$$\begin{array}{l}
\mathit{deceived\_} : \mathbb{P}(\mathit{NormativeAgent} \times \mathit{NormativeAgent} \times \mathit{Norm} \times \mathit{EnvState}) \\
\hline
\forall\, ag_1, ag_2 : \mathit{NormativeAgent};\ n : \mathit{Norm};\ nmas : \mathit{NormativeMAS}; \\
\quad st : \mathit{EnvState} \bullet \mathit{deceived}\ (ag_1, ag_2, n, st) \Leftrightarrow \\
\qquad (\ ag_1 \in n.\mathit{beneficiaries} \land\ ag_2 \in n.\mathit{addressees} \land \lnot\ \mathit{fulfilled}\ (n, st))
\end{array}$$

In addition, defenders also have a different relation with addressees. When a norm is activated, defenders are entitled only to enforce a norm, but when the norm is violated they have the responsibility to start a sequence of events leading to punish rebellious agents. Nevertheless, it could be possible that none of the defenders realises the occurrence of these events, and consequently the rebellious agent never becomes punished. Then it can be said that an agent must *punish* another agent if the first is a defender of the norm and the second is an agent who has violated it.

$$\begin{array}{l}
\mathit{punishes\_} : \mathbb{P}(\mathit{NormativeAgent} \times \mathit{NormativeAgent} \times \mathit{Norm} \times \mathit{EnvState} \\
\qquad\qquad \times \mathit{NormativeMAS}) \\
\hline
\forall\, ag_1, ag_2 : \mathit{NormativeAgent};\ n : \mathit{Norm};\ nmas : \mathit{NormativeMAS}; \\
\quad st : \mathit{EnvState} \bullet \mathit{punishes}\ (ag_1, ag_2, n, st, nmas) \Leftrightarrow \\
\qquad (\mathit{isdefender}\ (ag_1, n, nmas) \land ag_2 \in nmas.\mathit{members} \land \\
\qquad\quad ag_2 \in n.\mathit{addressees} \land \lnot\ \mathit{fulfilled}\ (n, st))
\end{array}$$

As we can observe, all these relationships are relativised both to a normative multi-agent system to which agents belong, and to the prevailing situation of agents. That is, no relationships due to norms can be created when agents do not belong to the same system, or when the conditions to activate a norm do not hold. We say that in a normative multi-agent system where social control has been defined through norms, some relations can be identified. That is, at a particular time, there are responsibilities that agents acquire through norms, situations in which addressee agents can be excepted from such responsibilities, enforcement mechanisms that might be applied to agents with duties, rewards that must be given to respectful agents, punishments that must be applied to norm offenders, and deceived agents expecting compensations. All these relationships change as soon as new norms become activated, fulfilled or violated.

6 Conclusion

So far in our work, the basic components of a system controlled by norms have been identified. We call these kinds of systems *normative multi-agent systems,* and we describe them as consisting of: a set of member agents whose compliance with norms is neither always enforced nor always expected, a set of norms directed at controlling the behaviour of all members, a set of legal norms to enforce compliance with regulations through punishment, a set of legal norms to reward agents who fulfill norms, and a set of norms to entitle some agents to change regulations. In general, current models of multi-agent systems regulated by norms include norms as obligatory actions that might otherwise be penalised [6,9,11, 16]. They typically do not make any distinction among norms. By contrast, our model divides these norms into three different classes which allow agents not only to identify the roles of other agents in a society but also to identify the limits of their responsibilities (given by the normative goals of the norms they have to comply with). In this way, agents' authority can be constrained.

Our model of norms facilitates the modelling of norms that must be complied with depending on compliance with other related norms (*contrary-to-duty* norms). By using interlocking norms, mechanisms to enforce compliance with norms are given through enforcement and reward norms. In addition, the *dynamism* that occurs in a system due to norms has been analysed and, according to the different stages in the processing of norms, some *normative relationships* have been identified. The key concept here is the normative behaviour of agents caused not only by the existence of norms, but also by their issuance, fulfillment or violation, which in turn must be the result of the decisions of each of the members.

By studying the characteristics of normative multi-agent systems, we have set up the basis of a framework to represent different kinds of social systems regulated by norms that include elements that allow agents to reason about norms. In addition, the set of normative relationships identified in this paper might enable agents to take more effective decisions in situations where norms are involved. For example, agents who have benefited from a fulfilled norm might decide to reciprocate with the addressees of such a norm in their subsequent interactions. Normative relationships are also useful for identifying situations in which a subset of agents is legally empowered, and informing about the decision of when a new norm can be adopted or complied with. This is the focus of the next stage in our work. We also aim to extend our work on norm compliance [14] to introduce strategies in which agents are externally influenced to comply with a norm. Additionally, we must provide an analysis of those situations in which agents might adopt new norms. We believe that the normative roles that we have defined here can be used by agents to identify empowered agents, and therefore to identify from whom an order can be received.

Our analysis builds on much important work on norms. Ross, for example [17], describes some of the norms and relationships presented in this paper. In turn, Conte and Castelfranchi [3] have already mentioned some of the normative roles we present, and some of the processes involved in reasoning about norms. Jones and Sergot [12] also mention the characteristics of agents entitled to manage an institution. The closest work is by Balzer and Toumela [1], who present the formalisation of an institution controlled by norms. However, no work considers the dynamics of norms nor the relationships that emerge from them, for use by agents to reason about norms.

Acknowledgments. The first author is supported by the Faculty Enhancement Program (PROMEP) of the Mexican Ministry of Public Education(SEP) and the Benemérita Universidad Autónoma de Puebla (México). Authors also express their gratitude to the valuable comments provided by the anonymous reviewers of this paper.

References

1. W. Balzer and R. Tuomela. Social institutions, norms and practices. In C. Dellarocas and R. Conte, editors, *Social Order in Multi-Agent Systems,* pages 161–180. Kluwer Academic Publishers, 2001.
2. G. Boella and L. Lesmo. Deliberative normative agents. In C. Dellarocas and R. Conte, editors, *Social Order in Multi-Agent Systems,* pages 85–110. Kluwer Academic Publishers, 2001.
3. R. Conte and C. Castelfranchi. *Cognitive and Social Action.* UCL Press, 1995.
4. R. Conte and C. Castelfranchi. Norms as mental objects. From normative beliefs to normative goals. In C. Castelfranchi and J. P. Müller, editors, *From Reaction to Cognition (MAAMAW'93),* LNAI 957, pages 186–196. Springer-Verlag, 1995.
5. R. Conte, R. Falcone, and G. Sartor. Agents and norms: How to fill the gap? *Artificial Intelligence and Law,* 7(1):1–15, 1999.
6. C. Dellarocas and M. Klein. Contractual agent societies: Negotiated shared context and social control in open multi-agent systems. In C. Dellarocas and R. Conte, editors, *Social Order in Multi-Agent Systems,* pages 113–133. Kluwer Academic Publishers, 2001.
7. F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law,* 7(1):69–79, 1999.
8. F. Dignum, D. Morley, E. Sonenberg, and L. Cavendon. Towards socially sophisticated BDI agents. In E. Durfee, editor, *Proceedings on the Fourth International*

Conference on Multi-Agent Systems (ICMAS'00), pages 111–118. IEEE Computer Society, 2000.

9. V. Dignum and F. Dignum. Modelling agent societies: Coordination frameworks and institutions. In P. Brazdil and A. Jorge, editors, *Progress in Artificial Intelligence Knowledge Extraction, Multi-agent Systems, Logic Programming, and Constraint Solving,* LNAI 2258, pages 191–204. Springer-Verlag, 2001.
10. M. d'Inverno and M. Luck. *Understanding Agent Systems.* Springer-Verlag, 2001.
11. M. Esteva, J. Padget, and C. Sierra. Formalizing a language for institutions and norms. In J. Meyer and M. Tambe, editors, *Intelligent Agents VIII (ATAL'01),* LNAI 2333, pages 348–366. Springer-Verlag, 2001.
12. A. Jones and M. Sergot. A formal characterisation of institutionalised power. *Logic Journal of the IGPL,* 4(3):429–445, 1996.
13. F. López y López and M. Luck. Modelling norms for autonomous agents. In *Proceedings of the Fourth Mexican International Conference on Computer Science (ENC'03) (to appear).* IEEE Computer Society Press, 2003.
14. F. López y López, M. Luck, and M. d'Inverno. Constraining autonomy through norms. In C. Castelfranchi and W. Johnson, editors, *Proceedings of The First International Joint Conference on Autonomous Agents and Multi Agent Systems AAMAS'02,* pages 674–681. ACM Press, 2002.
15. M. Luck and M. d'Inverno. A formal framework for agency and autonomy. In V. Lesser and L. Gasser, editors, *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95),* pages 254–260. AAAI Press/MIT Press, 1995.
16. Y. Moses and M. Tennenholtz. Artificial social systems. Technical report CS90-12, Weizmann Institute, Israel, 1990.
17. A. Ross. *Directives and Norms.* Routledge and Kegan Paul Ltd., 1968.
18. M. Sergot. Normative positions. In P. Mc Namara and H. Prakken, editors, *Norms, Logics and Information Systems,* pages 289–308. IOS Press, 1999.
19. M. Singh. An ontology for commitments in multiagent systems: Toward a unification of normative concepts. *Artificial Intelligence and Law,* 7(1):97–113, 1999.
20. J. M. Spivey. *The Z Notation: A Reference Manual.* Prentice-Hall, 1992.
21. R. Tuomela. *The Importance of Us: A Philosophical Study of Basic Social Norms.* Stanford University Press, 1995.
22. A. Walker and M. Wooldridge. Understanding the emergence of conventions in multi-agent systems. In V. Lesser and L. Gasser, editors, *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95),* pages 384–389. AAAI Press/MIT Press, 1995.
23. R. Wieringa, F. Dignum, J. Meyer, and R. Kuiper. A modal approach to intentions, commitments and obligations: Intention plus commitment yields obligation. In M. Brown and J. Carmo, editors, *Deontic Logic, Agency and Normative Systems,* pages 80–97. Springer-Verlag, 1996.

Integration of Generic Motivations in Social Hybrid Agents

Fenintsoa Andriamasinoro and Remy Courdier

IREMIA, University of La Réunion, BP 7151, Messag Cedex,
97715 Saint-Denis de La Réunion, France
{fenintsoa.andriamasinoro, remy.courdier}@univ-reunion.fr

Abstract. Most hybrid agent architectures are constructed with a hierarchical succession of reactive (at a lower level) and cognitive (at a higher level) layers. Each of these layers represents a behavior, a function, a decision, etc. Instead of using such functional layers, we propose in this paper a generic model of a social hybrid agent, which is based on natural (human/animal) motivations of the agent. We discuss here the contribution of our approach in hybrid agent modeling. The present work uses the American psychologist Abraham Maslow's pyramid of needs. The basis of this modeling uses the result of an existing psychological study.

1 Introduction

1.1 Human and Animal Behavior

In agent modeling, animals are considered as reactive entities, behaving according to their internal and external impulse and the dynamics of the environment, either in an individual or in a social context [4,8]. On the other hand, humans are regarded as cognitive entities which can evaluate the actions to be performed (we call this a high-level behavior), according to many parameters "imposed" by their society (policy, role, etc.) [10]. This high-level behavior in a human being is the realization of its animal instinct performed in a more rational form (intentional coordination, etc.). For example, on the one hand, when a hungry dog finds food in a kitchen, it immediately eats it (we call this a low-level behavior). On the other hand, a person first asks to whom the food belongs, and if he may eat it, its social norm leads it to first take a plate and a fork, etc. In any case, this person's behavior aims to satisfy the animal instinct in him, which is the hunger. The combination of both behavior levels forms the hybridism.

Besides the social norm, the human also makes, for instance, a high-level organization: the need for eating is manifested at a higher-level by the need for everybody to work. The remuneration resulting from this work will then be used for buying foods, satisfying then the need to eat (where *eating()* is the low-level behavior).

We may see from the above concepts that humans and animals have common basic motivations: the satisfaction of natural needs (hunger, sexual impulse, etc.) but the two categories differ in the way they satisfy them. This concept is also

G. Lindemann et al. (Eds.): RASTA 2002, LNAI 2934, pp. 281–300, 2004.

used in agent modeling, [2,4]. These motivations (also called source of actions by [6])make an agent behave either reactively or cognitively (i.e. adopting a low or a high level behavior). As we see, the concept of *motivation* takes an important place in the study of agent behavior.

The objective of the work presented in this paper is to integrate this theory of basic motivations in hybrid agent paradigm, by taking into account these two levels of behavior. The idea here is to propose a generic model that integrates abstract motivations (also called "abstract common source" of actions) which may then be instantiated according to the studied application. For this purpose, we use Abraham Maslow's pyramid of needs [1,2, 13].The basis of this modeling thus uses the result of an existing psychological study.

Because we model the human/animal behavior and we also use the pyramid of Maslow (initially based on human needs), our work then concerns the modeling of social agents, those having the need to live in a society [4,10]. This work integrates the concept of hybrid agents [19] that will have needs, feelings[1], and so on, during their activities. It is obvious that human and agent have differences. For example, agent is a computer-generated entity and then, is "physically" limited by the computer capabilities while human is a nature creature, having a larger life dimension and possibilities (which theoretically tend to the infinite). But with regards to our current purpose, the general human behavior can be modeled.

The terms *needs* and *motivations* presented in this paper may be confusing. Actually, the basic motivation of an agent corresponds to the satisfaction of its (or others) basic needs (see Section 3.2). However, as they are in fact equivalent, we adopt the idea that these two terms can be alternatively used, depending on the context of our explanation.

1.2 Modeling Motivations in Hybrid Agents: State of Art

Hybrid agent architectures. Most hybrid architectures are constructed of layers, each of them defining a specific function and possibly a decision. We may mention TOURINGMACHINES [7] a model having three layers (from bottom to top): the reactive, the planning and the modeling layer. Another layered architecture is INTERRAP [15]. In this approach, each successive layer represents components: a behavior-based, a plan-based and a cooperation-based component, the overall is connected in a knowledge base. More recent layered architectures are ICAGENT [14] which models the intention reconciliation and planning in agents, and GLA [12] in which layers regroup similar types of computations instead of a functional decomposition as in the previous models.

The common characteristic of these layered architectures is that the layer components generally define behavior, plans, cooperation, decision, etc. We also note that the lowest layer of most of them integrates the reactive part of the model (because it is close to the environment), followed at a higher layer level by the cognitive part (and possibly other additional layers).

[1] The notion of feelings will be studied in a future work but the present paper outlines our idea for doing it.

Relation between motivation and actions. The layers in the architectures presented above model actions rather than their source: the agent motivations [6]. If we consider the hybridism concept from the angle of a "combination" of reactive and cognitive models, the notion of motivations may be found, particularly at the reactive level. At this level, the motivation (satisfying hunger, avoiding obstacles, etc.) is the main factor that determines the behavior of the agents, followed by an action selection process [8]. In this case, we consider the motivation as explicit. The above examples are however chosen depending on the application (ants, robots, etc.).

On the other hand, in works about cognitive modeling, the motivation concept is less considered even if the dynamic of the environment (leading to a reactive/instinctive behavior) is taken into account [3,18]. Such works are more focused on organization, coordination, etc. However, some of them use the notion of *desire,* particularly those using the BDI framework [5], in which a *desire* is a goal driven by a motivation. But even in these cases, the notion of motivations is not very clear and finally, considered as abstracted by the agent designer.

1.3 Objective and Issues

In brief, the modeling of motivation in a hybrid agent is either not considered at all or considered but only depending on application. In this paper, we propose a generic hybrid agent model called MASLOW (acronym for Multi-Agent System based on LOW needs) from the same name as the psychologist, but also especially based on Low-Needs concepts (Section 3.3). By integrating this generic model, we aim to overcome what we consider to be a "lack of motivation" concept in hybrid agent modeling.

The problem related to this work is the balancing between the reactive and cognitive agent behavior when it is known that an agent has a goal-directed behavior when it has to cope with the environments[2] (e.g. [11]). In addition, and given that we base our work on motivations, another issue is to determine the degree of motivations for the agent for satisfying each need of the pyramid. The one having the higher degree is treated first. In our current study, this problem is oriented towards the determination of the need semantically being the most important. Criteria must be given. Obviously, the importance of needs may vary from time to time and then, a repetitive process of checking this importance must be performed.

1.4 Preamble

In our work, the cognitive part of the agents is not as yet studied deeply. However, we plan to integrate it in the future (see Section 9.2). Meanwhile, many assumptions are first made to deal temporarily with this situation. We assume that cognitive agents already have a plan (see Section 5.1 about this notion) and

[2] The environment of the agent may be the internal one, such as impulse, etc., or physical one or the other agents in the system

each cognitive agent has knowledge of the need state of others. We agree that in case, this latter assumption is not realistic because in a general way, the knowledge of an agent about its environment is partial [6,9]. Thus, at this stage of the work, it may represent a limit of our model and requires further investigation. However, it does not really affect the modeling of our hybridism concept, that is, the balancing between the reactive and the cognitive behavior of our agents (Section 5.2).

MASLOW is currently developed in JAVA. Thus, some notations in this paper also follow the syntax of this language.

The remainder of this paper is organized as follows: Section 2 presents an overview of the case-study we analyze throughout this paper, followed by the basic concept of our model: the description of needs (Section 3). Next, Section 4 particularly presents our concept of social needs according to both the pyramid definition and the agent modeling. After presenting these concepts of needs, we outline in Section 5 their relation to actions. All of these parameters are necessary before we can define in Section 6 the criteria needed in the management of the importance of the need. The whole model is evaluated in Section 7 and analyzed in Section 8. Lastly, Section 9 concludes the paper and gives our perspectives of the work.

2 Case Study

Our current case study concerns two docker agents D_1 and D_2 working in a harbor and paid daily. The wage from the work will be used to buy foods. The aim of the job is to carry heavy goods from a storage to a boat (storage→ boat=l journey). In the storage, there are many goods but the daily work consists of carrying just 8 of them (only 1 unit per journey is possible). Additionally, the dockers have each a bottle of water so that during the convey, they also can drink. The adopted coordination made by the docker association is that D_1 will carry 3 goods whereas D_2, 5 goods, both following a road, and assuming that at the road's edges, there are "dangerous ravines" beyond which, there is the sea (Figure 1). As the harbor is a dynamic environment (ex: containing grounded obstacles, a crane which may inadvertently "release" something, etc.), D_1 and D_2 must consider this dynamic when performing their work.

After the docker carriage repartition (respectively 3 and 5 goods) is made, each of them commits himself to fulfilling his respective task because they want to show that they do a good job. By doing so, they want to be integrated and to be respected, either by each other, or by the association itself. Additionally, the first docker who will have finished its job may help the other one.

The initial cognitive plan P of the docker is: *P:={goTo<storage>, take<good>, goTo<boat>, put<good>}* until the 8 (=3+5) ores are carried. This plan is then the high-level behavior of the dockers.

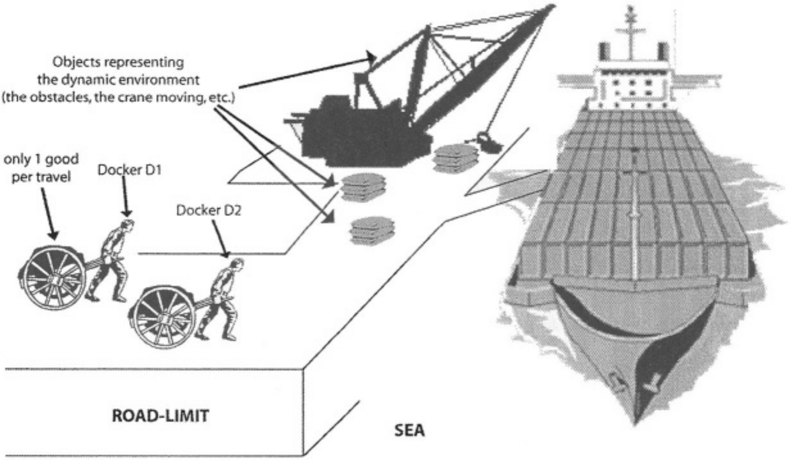


Fig. 1. The scheme of the case study. The two dockers D_1 and D_2 have the cognitive plan to carry goods but in a dynamic and dangerous environment.

3 Basic Concepts

3.1 The Pyramid of Abraham Maslow

Our present work uses the Pyramid of MASLOW (hence noted Π) of the American psychologist Abraham Maslow [13]. Π regroups the five hierarchical needs of a Human Being: physiological needs, the need for security, the need for love, the need for esteem and the need for self-realization. This last level is not yet analyzed deeply in this work.

According to this psychological study, *all actions led in the Living Being's behavior are motivated by at least one of these five hierarchical needs.* We call them *abstract* needs as each species of Living Beings has its own (or sometimes common) way to satisfy them. But the final objective is the same: to satisfy one or another of these basic needs. As the needs are abstract, terms such as need for *security, social,* etc. actually corresponds to any needs which are set at these levels, which are then not always unique.

The architecture of the pyramid itself is one of our reasons for choosing this model in our work (additional reasons are found in [2]). Indeed, it results from a psychological study of human behavior (like biologists study animal behavior). In our agent modeling, each level then has a specific conceptual semantic not based on pure hypothesis.

3.2 The Pyramidal Need

Formalization. The fine-grain need of Π is called PN for Pyramidal Need. A formalization of PN was presented in [2]. Generally, it corresponds to the need state: the *quiet (sufficient),* the *threatened (limit)* and the *missing (insufficient)*

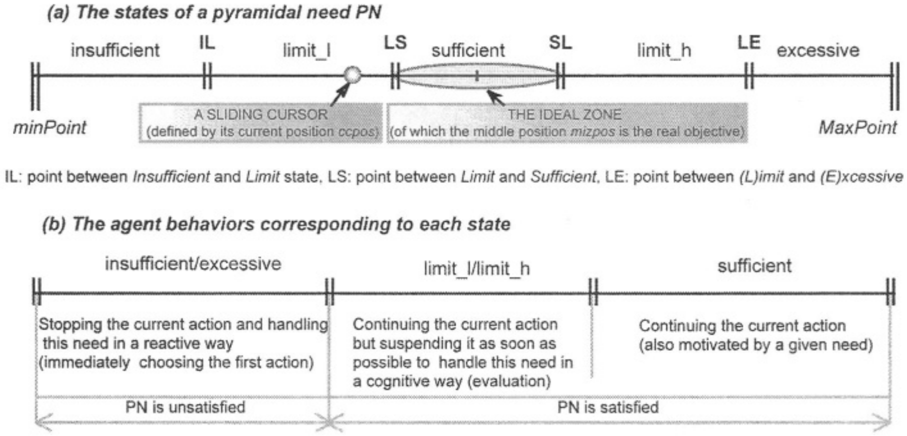

(a) The states of a pyramidal need PN
insufficient
IL
limit_l
LS
sufficient
SL
limit_h
LE
excessive
minPoint
A SLIDING CURSOR (defined by its current position ccpos)
THE IDEAL ZONE (of which the middle position mizpos is the real objective)
MaxPoint
IL: point between Insufficient and Limit state, LS: point between Limit and Sufficient, LE: point between (L)imit and (E)xcessive
(b) The agent behaviors corresponding to each state
insufficient/excessive
limit_l/limit_h
sufficient
Stopping the current action and handling this need in a reactive way (immediately choosing the first action)
Continuing the current action but suspending it as soon as possible to handle this need in a cognitive way (evaluation)
Continuing the current action (also motivated by a given need)
PN is unsatisfied
PN is satisfied

Fig. 2. The formalization of a Pyramidal Need (PN)

ones. However, this formalization does not take into account the state where the need is “over satisfied” (in an excessive way) while it actually may occur in many real situations. An example of this last state is a person who is fed too much (the need ‘hunger’ is over satisfied) or who has high blood pressure, etc. We then first introduce this state in this paper (see Figure 2.a). thus, we also “split” the limit *state* to “*limit_l* ” (“1” for “*low* ”) and “limit_h” (“h” for “*high*”).

A PN is normally written $PN_{level/rank}$ in which *level* and *rank* are the indexes (like a “physical” place) of the need in Π. But in our formalization, these two parameters may be omitted if not necessary (just write PN). Note that the *level* parameter is considered during the determination of the degree of motivation of a PN while the *rank* one is not (see Section 6 for more precisions).

The states are presented throughout an axis on which a cursor - the indicator of the current state - slides. We call its current position *ccpos.* The zone corresponding to the *sufficient* state is called the *ideal zone* of which the middle position is called *mizpos* for middle ideal zone position. If *ccpos==mizpos,* the need is fully satisfied. Each PN need has a unit called *PNu* (e.g. liters, meters, pounds, etc. depending on the application).

A state is formalized as an interval as follow (‘//’ means a comment):
state=(born)minPoint, IL(sign), LS(sign), SL(sign), LE(sign), maxPoint(born)
in which
- *born* is either ‘]’ or’[‘ depending on the fact this side of the interval is closed or opened.
- *sign* is either ‘<’ or ‘>’. If we write ‘IL<’, it means that the *insufficient* state (situated on the left side of IL) is closed. It automatically involves that the *limit* state (on the right side) is opened // *IL, LS, SL, LE are the points which separates states* (Figure 2-a).

As example about D_2, the transport of the goods is formalized as follows: state=[0, 3<, 5>, 5<, 100>, 100].
Concretely, this example means that D_2 will not feel satisfaction until the 5 goods he is committed to carrying are carried. But D_2 also feels the same sensation if he thinks of carrying beyond 5 goods (the state of *work* is going to the *excessive* state), the above formalization meaning that there is a given reason for not doing so (e.g. no more wage even if conveying more goods, or, no more place in the boat to stock them, or, the boat cannot support more than 8 goods, etc.). The value 100 in the above example is a random chosen number that the need cannot reach (in fact, the representation of a PN currently constraints us to create a intervals with finite values only).

3.3 Low/High Needs (LN/HN)

The Low-Needs (LN) and High-Needs (HN) [2] are the only possible type of PN. As a need PN is either LN or HN, these two notations constitute the *type* of PN. In sum, $\{PN\}=\{LN\}\cup\{HN\}$.

On the one hand, the LN corresponds to the "inborn" needs of the agents, also called the *natural parts* of the agents: hunger, sleep, preservation instinct, etc. The LN is *permanent* and is always *active* (considered in behavior). On the other hand, the HN corresponds to the needs that (only) cognitive agents have, resulting from its plan, intention, reasoning, etc. HN is temporary and is active or not. Indeed, in general, high-level goals differ from one agent to another (even if in our case study, they are the same).

What we emphasize is that one *HN is always motivated by at least one LN* (see Equation(1)). Symmetrically, one LN may be the motivation of more than one HN. *LN* is an impulse to be satisfied and *HN* is the cognitive adopted goal (in form of desire) to directly or indirectly satisfy *LN*. The docker D_1 has for instance the desire *HN* with *HN.sufficient:= nbGoodsCarried==3* to be satisfied. But the satisfaction of this need is actually motivated by the satisfaction of a *LN:=satisfy_hunger,* that is, the wage from the work will be used to buy some food.

This relation between HN and LN is formalized via a functional relation $f \in F$(Equation (1)) joining the high and low level needs where f is actually an action and F designates the set of them. The function f may be first a composition of other functions. In other words, it may happen that $f = f_1 o ... o f_n$, *with* $\overline{1..n} \in \mathrm{N}$

$$\forall LN, HN \in \Pi, \exists f \in F / LN = f(HN) \quad (1)$$

Obviously, the inverse of Equation (1) is not always true (i.e. f is not surjective). The *LNs* do not always have a corresponding *HN* because an agent has not always to perform a high-level behavior in order to satisfy a low-level need.

Figure 3 summarizes our description of needs. All needs are managed by a module called the Need Importance Manager (NIM), which gives the present most important need requiring priority treatment (see Section 6).

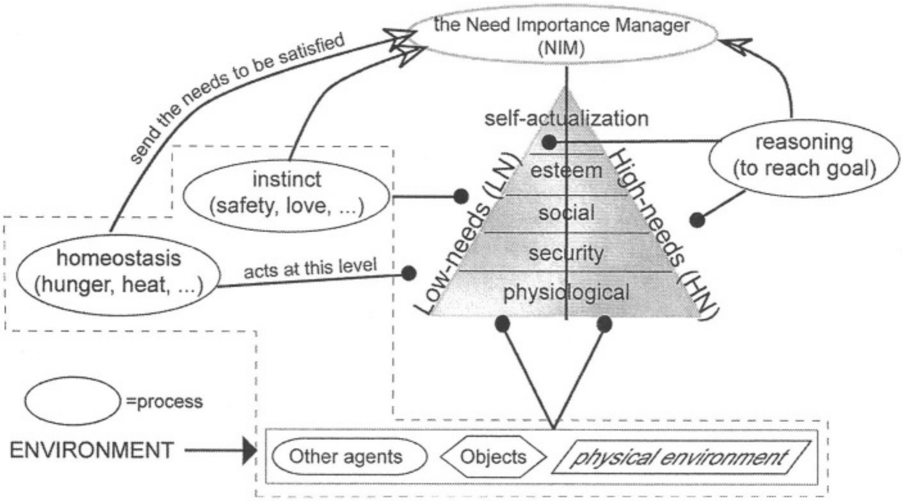


Fig. 3. The hybrid architecture based on need managements

The environment of an agent is composed by all components of the system which manage or modify the needs. According to this definition, the reasoning process is not then included in the environment. It is rather a "mental tool" that the agent uses to reach its goal.

3.4 The Need Variation Speed (PNVS)

The PNVS is the speed at which the cursor of a PN is sliding on the axis. It is an important parameter as it determines the approximate time for the agent to react. For instance in the case where an important PN is worsening, the more the PNVS is elevated, the more agent "feels" to quickly treat this PN. Note that the only source of a PN-cursor moving (if any) is internal or external actions (i.e. from the agent or from the system).

The PNVS is evaluated as follows: there is an initial time t_0 which is initialized, either at the first time the PN is created in the agent, or the last time the PNcursor in the axis has changed direction. Let x_0 the cursor-position at t_0. The value of PNVS is determined by Equation (2).

$$PNVS = \frac{ccpos - x_0}{|current\_time - t_0|} * \frac{ccpos - mizpos}{|ccpos - mizpos|}(PNu/timeunit)\,. \quad (2)$$

The operand on the right only aims to get the sign of the current direction of the cursor, compared to the ideal zone. Then, if the resulting PNVS<0 (respectively>0, or ==0) then the state of PN is said *worsening* (respectively *improving,* or *stationary*).

3.5 Notations and Definition

For the best comprehension of the remainder of the paper, the following notations must be set:

- predicates *PN.isxxx()* indicates that the description attributed to *PN.xxx* is verified, *xxx* being either one of the need state (e.g. *PN.isSufficient()*), or the evolution of the need state (e.g. *PN.is Worsening()*).
- the following relation is mentioned:
  insufficient/excessive < limit < sufficient
  (at the present period of our research, no comparison can be made between the *insufficient* and *excessive* state).
- predicates *isHN(PN)* and *isLN(PN)* respectively indicates if PN is a HN or a LN (see Section 3.3). Additionally, there is a function named *type_of(PN)* which returns the type of PN (LN or HN). Then, *isHN(PN)==true⇔type_of(PN)=HN.*
- when a need PN_1 is *more important* than another, PN_2, it is noted $PN_1 > PN_2$ (and meaning that PN_1 must be treated before PN_2).

In addition, in this paper, we call *the need checkpoint* the moment during which the Need Importance manager or NIM (see Figure 3) checks the state of all other needs in the pyramid and determines if there is or is not a more important need (than the current being treated) to be satisfied. Section 6 explains the way in which an important need is found.

4 The Levels 3 and 4: The Social Needs

4.1 Principles

Like the individual needs, the social ones of Maslow (the abstract needs to be liked or to be esteemed) are also motivations of agents' behavior in social context.

We note the following basic needs:

- not to be alone: this corresponds to the basic need to be in an environment where there is at least another congener (even if at this stage there is not yet any relationship between them).
- to be integrated or to be loved: this need is also valid for animals [4]. Example, the ones which integrate a group (to be liked), a natural *chief* of an animal society who wants to be respected (to be esteemed), etc.
- considering the others: particularly the impact of its action on others [10]. As far as possible, the agents generally do not take actions which worsen the need state of others.

At a HN level, many acts can be mentioned when looking for love, esteem, consideration, etc.:

- the need for commitment when working[3]: by satisfying this need, the dockers want to show to others that they work perfectly. They also know that if they do not do so, they will be "rejected[4]" by their society (in [6], it is called a *functional motivation).* At the current stage of our research, the concept of *being loved* is limited to only this functional motivation. But we agree that in a future, different "stages" of love must be considered: to be loved by family, by friend, loved by colleagues, etc.
- helping: helping is also a social act. But its realization depends on the other(s) agent(s) to be helped (congener? family? son? friend? etc.). The basic motivation for helping may then vary: the social obligation to working well (also a functional motivation), the search for a friendship in return for the help (=searching love), the search for esteem from the other, etc.

5 Needs and Actions

5.1 Formalization of Actions

There are two kinds of actions:

- a *primitive,* the fine-grain action. It is uninterruptible when executed. Due to this characteristic, the need checkpoint is possible at least only between the execution of two consecutive primitives.
- a *plan,* composed by one or more primitives and intentionally prepared by cognitive agents. To be executed, a plan is recursively decomposed like a tree, until having the leaf (the primitives). By definition, a sub-plan is a part of a plan but situated at a lower-level in the tree.

The relation between needs and actions is set as follows: when executed for satisfying a need, an action α is repeated until a condition, called a *local_need* is satisfied. It is a PN locally related to a given action. Each action then has its associated *local_need.* For example, a docker who is going to the storage has the following parameters:
- the action α= *Goto(storage),*
- having PN=local_need(α)/PN.sufficient:=*self.position==storage.position, // note that here, PN is a HN because associated to a plan*
it means that the docker has a local (and psychological) need to reach the storage (i.e. want to have the same position as it). He willperform α until this need is satisfied.

5.2 From Reactive to Cognitive Behaviors

To satisfy a given need, there is a list of n ($n \geq 1$) actions among which agents may choose. Choosing an action in a reactive way means that agent randomly takes

[3] We note that we consider here only the social level. But we agree that the commitment may also be made in the context of individual domain (self-commitment)

[4] At a basic level, being rejected is felt as not being loved any more.

one action between these n actions. Doing so in a cognitive way means that the agent evaluates each action. We thus use the Action Selection Mechanism (ASM) [8]. Our criteria during the action selection process are presented in [2]. The balancing between the reactive and cognitive behavior depends on the current state of the need PN to be treated. When the PN-state is *insufficient/excessive,* the agent acts reactively. Otherwise, it acts cognitively, and by evaluating the appropriate action.

Acting in order to treat an unsatisfied need PN means suspending the current action the agent is performing. The suspension is first planned when the need PN to be treated is in a *limit* state. Furthermore, the suspension of the current action (let β) (also performed due to a previous unsatisfied need PN) can be possible, only between two primitives. If so, the model checks if PN is more important than PN'. In such a case, PN will be treated. Otherwise, the satisfaction of PN' via β is resumed.

6 Managing the Need Importance

6.1 Recall

The choice of the next action depends on determining first which need is the most important at the phase of checkpoint (assume, at the moment t). The need checkpoint (previously defined in Section 3.5) then actually consists of determining, for the time t, the degree of motivation for the agent to satisfying each PN (noted PNDM) of his Π. As the value of these PNDM may vary from time to time, the result of the checking at time t is then valid for only this time. This means that a need may not be important at a time t but may be so at $t+1$, depending on the agent activity.

If the checking issue results that PN1>PN2 ($PN1,\ PN2 \in \Pi$), it means that the agent is more motivated to satisfy PN1 than PN2.

The degree of motivation for a PN depends on its type (HN or LN), his level in Π and his *state_ratio* (see Section 6.2, §*the urgency).* In other words, *PNDM=f(type_of(PN), level(PN), state–ratio(PN)).*
This relation derives from the criteria we already proposed in [2]. The relation between PNDM, the type and the level is determined by the criteria recalled in Equation (3), in which x, i, j represents a level number.

$$\begin{array}{l} 0-\forall PN_{i/rank},\ PN_{j/rank} \in \Pi \Rightarrow PN_{i/rank} > PN_{j/rank}\, \forall i < \underline{j}\ \ (i,\ j \in \overline{1..5}) \\ 1-\forall LN_{i/rank},\ HN_{i/rank} \in \Pi \Rightarrow LN_{i/rank} > HN_{i/rank}\ (i \in \overline{1..5}) \\ 2-if\, \exists LN_{i/rank},\ HN_{x/rank} \in \Pi,\ \exists f_i \in F/HN_{x/rank} = f_i(LN_{i/rank}) \Rightarrow x := i \end{array} \tag{3}$$

6.2 New Criteria

Resolution of a level classification problem. The problem in Equation (3) is that Criterion 1 only compares HN and LN at the same level. When they are

in a different one, it no longer holds, and it seems that Criterion 0 is better. However, if we strictly apply Criterion 0, the expression:
$\forall i < j, \in \overline{1..5}, \forall HN_{i/rank}, LN_{j/rank} \in \Pi, \Rightarrow HN_{i/rank} > LN_{j/rank}$
becomes true, and involves for instance that the high-need for transporting a good (a $\mathrm{HN}_{1/rank}$ motivated by the need to eat) is more important that avoiding an object falling down from the crane (a $\mathrm{LN}_{2/rank}$ need related to the security). Nevertheless, we know that this situation is not always realistic particularly when the need to be secure is threatened. Thus, in such a case, Criterion 0 may not be exceptionally applied. The new criterion described by Equation (4) is a proposed solution. It stipulates that if the state of a LN_i is *unsatisfied* (and only in this case) while the agent is performing a HN_j (with j<i), then the HN_j is suspended and LN_i is performed, *until it is led back to the limit state.*

$$\begin{array}{l} \forall x, i \in \overline{1...5}, \forall PN_{x/rank}, LN_{i/rank} \in \Pi \ / \\ \quad if\, LN_i.isUnsatisfied() \\ \quad\quad *\ if\, isHN(PN_{x/rank}) \Rightarrow LN_{i/rank} > PN_{x/rank} \\ \quad\quad *\ if\, isLN(PN_{x/rank}) \Rightarrow LN_{i/rank} > PN_{x/rank}\, (but\, only\, if\ x > i) \\ \quad otherwise\, valid(criterion\, 0) \end{array} \tag{4}$$

Cloning. Another principle is also introduced in this paper: cloning (see Equation(5)). In fact, it happens that one HN is motivated by more than one LN. Then, the HN is cloned as many times as the number of the corresponding needs LN. The interest of cloning is in the different "basic satisfaction" in which one given HN is involved. All cloned HN have the same structure but, once created, they then evolve differently.

$$\begin{array}{l} if\, \exists LN_1, ..., LN_k, HN_x \in \Pi, \\ \quad \exists f_1, ..., f_k / HN_x = f_1(LN_1), ..., HN_x = f_k(LN_k) \\ \Rightarrow create(HN_{x1}) := clone(HN_x), ..., \ create(HN_{xk}) := clone(HN_x) \\ \wedge HN_{x1} = f_1(LN_1), \ldots, HN_{xk} = f_k(LN_k) \end{array} \tag{5}$$

The execution of the plan P is for example based on three basic motivations: the motivation for having something to eat, the functional motivation and the motivation to be esteemed (as mentioned in Section 3.4). Figure 4 shows the principle of cloning (the right side of the pyramid) but also summarizes the general modeling of needs according to the scenario described in Section 2 and the relation in Equation (1). In Figure 4, when a LN has no corresponding HN, it means that its satisfaction does not depend on the current cognitive goal.

Classification of the local needs. This is made as follows:

- let α, β two plans, α being a sub-plan of β, if(HN_α=local_need(α) and HN_β=local_need(β))$\Rightarrow HN_\beta > HN_\alpha$
- if a plan α is made by a series of actions ($\alpha_1, \ldots, \alpha_i$) $\Rightarrow$ local_need(α_1) > $\ldots$ >local_need(α_i)

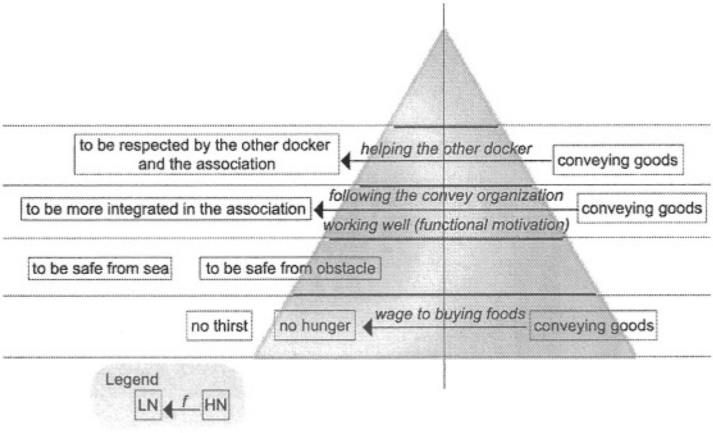

to be respected by the other docker and the association
helping the other docker
conveying goods
to be more integrated in the association
following the convey organization
conveying goods
working well (functional motivation)
to be safe from sea
to be safe from obstacle
no thirst
no hunger
wage to buying foods
conveying goods
Legend
LN
f
HN

Fig. 4. The principle of cloning.

Urgency. The notion of *urgency* is an additional criterion we first propose in this paper to resolve the case where, after applying both type and levels criteria, there are still two or more PN that have exactly the same importance. In fact, in [2], the state_ratio[5] noted *sr* was already a first solution for this case, but *sr* is only a spatial criterion. As the need is temporally dynamic, a spatiotemporal criterion is better. For that, we then use the PNVS proposed in Section 3.4.

Assume that, the preserving instinct of D_1 leads him to have two motivations:

1. to avoid a grounded obstacle situated at 2,5 *lengthunit* from him (=a need PN_1 to be away from the closer grounded obstacles). The cursor moving characterizes the moving of D_1, and the *minpoint* of PN_1 corresponds to the position of the grounded obstacle. Here, the PN_1VS is the speed at which D_1 is walking.
2. to escape from an "aerial" object which is falling down from the crane, and going to fall directly onto him (=need PN_2 to be away from "aerial" obstacles). The currently falling object is situated at 8 *lengthunit* above D_1. The *minpoint* of PN_2 is the current position of D_1. The PN_2VS is the speed of the heaviness corresponding to the force of the gravity.

According to only the *sr* criterion, the most important need to be treated will be PN_1 (because 2,5 lu<8 lu). But when we take into account the PNVS, the situation is somewhat different. Indeed, it is sufficient that 2,5*PN2VS > 8*PN1VS so that PN2 theoretically reaches its *minpoint*before PN_1 and in this case, the theory of *sr* is no longer valid. Actually, *sr* Criterion may be applied only if after the applying the urgencycriterion, PN_1 and PN_2 still have the same importance. On the whole,
PNDM=f(type_of(PN), level(PN), PNVS, state–ratio(PN)).

[5] As a reminder, the *state_ratio* is the current relative position of the cursor in the axis, compared to the *minPoint* (see Figure 2 about the *minpoint*)

To summarize, given i needs PN1, ..., PNi,:

1. applying first *PNDM=* f*(type_of(PN), level(PN))* as described in previous Subsections.
2. for the remaining PN needs (if any), comparing them by applying the principle of *urgency.*
3. about the remaining PNs (also if any), considering the current state of each of them
4. last, using the *state_ratio.*

7 Evaluation

7.1 Implementation

Our model is evaluated via a prototype. The implementation is organized in three main layers:

- a *kernel* layer, gathered under the *maslow.kernel* package and implements the classes of the concepts which are studied in the present work;
- an *appli* layer, found under the *maslow.appli.docker* package and corresponds top the implementation of the prototype;
- a *gui* (Graphic User Interface) layer. This last one is not actually an interface designed for the MASLOW model but is rather an adaptation of our generic GUI. This GUI is connected to MASLOW via the package *maslow.appli.docker.gui.* The connection is made at this level because we aim to build the structure of the kernel independently of any GUI, making it more flexible.

Each agent implements the interface *java.lang.runnable* (to respect its autonomy), and the method *run()* starts the behavior of the agent. The proactivity of the agent is driven by the satisfaction of needs according to their importance. For that, an agent balances between the checking and the satisfaction of a need via actions. The implementation sequence is shown in Figure 5.

7.2 Instantiation

Before evaluation, the needs of each docker must be first instantiated. We show below that of D_1.

The individual aspects

- $PN_{1/1}$: hunger⇒physiological (LN) // *when this need is in an insufficient state,* D_1 *can no longer move.*

- $PN_{1/2}=f_1(PN_{1/1})$: transporting 3 ores⇒physiological (HN) // *the local need of the plan P. This HN is situated at level 1 because the docker knows that one of his basic motivations for doing the work is to get something to eat (via the money of the wage).*

- $PN_{1/3}=f_2(PN_{1/1})$: going to the storage⇒ physiological (HN) // *the local need for reaching the storage location. It is situated at this level because it is the local need of a sub-action contributing to the realization of P (with $PN_{1/4}$) at physiological level.*
- $PN_{1/4}=f_3(PN_{1/1})$: moving to the boat ⇒ physiological (HN) // *like $PN_{1/3}$ but concerning the boat location*

- $PN_{2/1}$: safe from any grounded obstacles ⇒security (LN) // *avoiding any obstacles situated in the ground*
- $PN_{2/2}$: safe from any aerial obstacles⇒security (LN) // *avoiding any objects, e.g. that of falling from the crane.*
- $PN_{2/3}$: safe from any dangerous regions ⇒security (LN) // *regions are here the sea and the road-limit.*

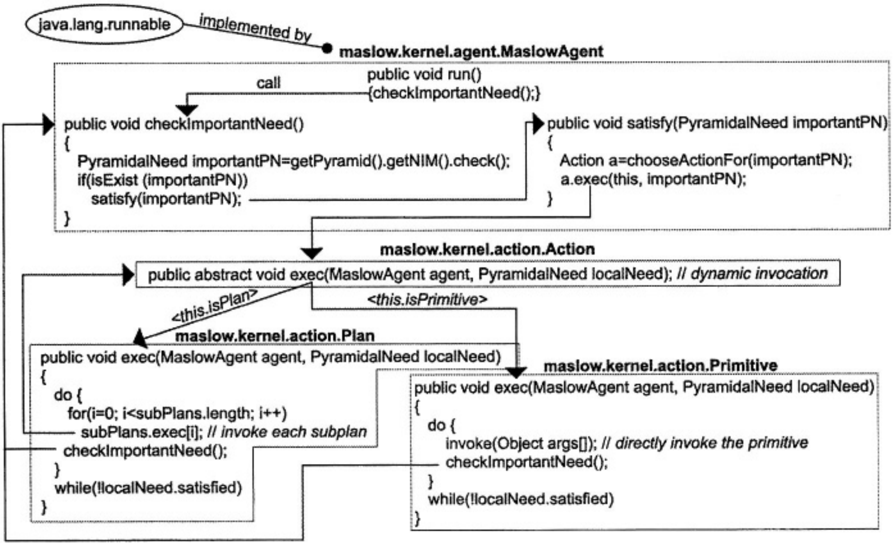
The social aspect —

- $PN_{3/1}$: not to be alone⇒social (LN) //*just knowing or seeing that there is a congener around him (intrinsic characteristic of social entities)*
- $PN_{3/2}$: to be integrated or loved⇒social (LN) // *being accepted in a group. Remind that the group may also just be a set of animals, not always a cognitive structure having a common goal. We are here at a LN level.*
- $PN_{3/3}$: consideration⇒social (LN) // *respecting the needs in the pyramid of the other*
- $PN_{3/4}$:=clone($PN_{1/2}$)=$f_4(PN_{3/2})$: transporting 3 ores⇒ social (HN) // *the local need of the plan P. This HN is a clone for this level 3 because the docker knows that one of his basic social motivations in doing the work perfectly is not to being rejected (it is a functional motivation as mentioned in Section 3.4).*
- $PN_{3/5}=f_5(PN_{3/2})$: going to the storage⇒social (HN) // *the local need for reaching the storage location. It is situated at this level because it is the local need of a sub-action contributing to the realization of P (with $PN_{3/4}$) at social level.*
- $PN_{3/6}=f_6(PN_{3/2})$: moving to the boat ⇒ social (HN) // *like $PN_{3/5}$ but concerning the boat location*
- $PN_{3/7}=f_7(PN_{3/3})$: helping the others⇒social (HN) // *helping the other during the work (if possible)*

- $PN_{4/1}$: to be appreciated ⇒esteem (LN)
- $PN_{4/2}$:=clone($PN_{3/7}$)=$f_8(PN_{4/1})$: helping the others⇒esteem (HN) // *helping the other during the work (if possible). The motivation is to be considered as a "good person" by D_2*

7.3 Scenario Results

According to the above scenario, and regardless of D_2, we present here some behavior of D_1, when the two dockers are carrying goods:


java.lang.runnable
implemented by
maslow.kernel.agent.MaslowAgent
call
maslow.kernel.action.Action
<this.isPlan>
<this.isPrimitive>
maslow.kernel.action.Plan
maslow.kernel.action.Primitive

```
public void run()
{checkImportantNeed();}
```

```
public void checkImportantNeed()
{
  PyramidalNeed importantPN=getPyramid().getNIM().check();
  if(isExist (importantPN))
    satisfy(importantPN);
}
```

```
public void satisfy(PyramidalNeed importantPN)
{
  Action a=chooseActionFor(importantPN);
  a.exec(this, importantPN);
}
```

```
public abstract void exec(MaslowAgent agent, PyramidalNeed localNeed); // dynamic invocation
```

```
public void exec(MaslowAgent agent, PyramidalNeed localNeed)
{
  do {
    for(i=0; i<subPlans.length; i++)
      subPlans.exec[i]; // invoke each subplan
    checkImportantNeed();
  }
  while(!localNeed.satisfied)
}
```

```
public void exec(MaslowAgent agent, PyramidalNeed localNeed)
{
  do {
    invoke(Object args[]); // directly invoke the primitive
    checkImportantNeed();
  }
  while(!localNeed.satisfied)
}
```

Fig. 5. The implementation of the proactivity of the agent: mixing the satisfaction of the current need and the checking of another important one

- D_1 avoids D_2 when they meet each other. It is due not only to their respective preservation instinct (when they meet, $PN_{2/1}$*.isLimit()*) but also due to the consideration of the state of the needs of the other ($PN_{3/2}$*.isLimit()*)
- likewise, the robots also avoid other obstacles (also due to $PN_{2/1}$). When an unexpected obstacle arrives close to an agent (for instance fallen down from the crane), $PN_{2/1}$*.isInsufficient()* is true. Then, the current plan P (motivated by $PN_{1/1}$) is immediately suspended (because of the criterion in Equation (4)) and an action among that corresponding to $PN_{2/1}$ is reactively chosen. The behavior is reactive because $PN_{2/1}$ is at the *insufficient* state. After this back tracking, the plan P is resumed to fulfill $PN_{1/2}$.
- in this scenario, it is noted that when D_1 has finished his own job (he has carried his 3 goods) while D_2 has not, D_1 will help D_2 in his work, because he knows that by doing so, he will improve the state of one of the needs of D_2 (satisfaction of $PN_{3/7}$, is motivated by $PN_{3/3}$).
- each time the water level of D_1 decreases, he drinks, leading this level to a better state again (illustrated in Figure 6).
- ...

8 Synthetical Analysis

Whilst the concept of motivations is really considered in reactive agents, works about cognitive agents focus their study on *how* an action will be realized (planning, reasoning, modeling of mental attitudes, etc.) rather than *why* (i.e. the

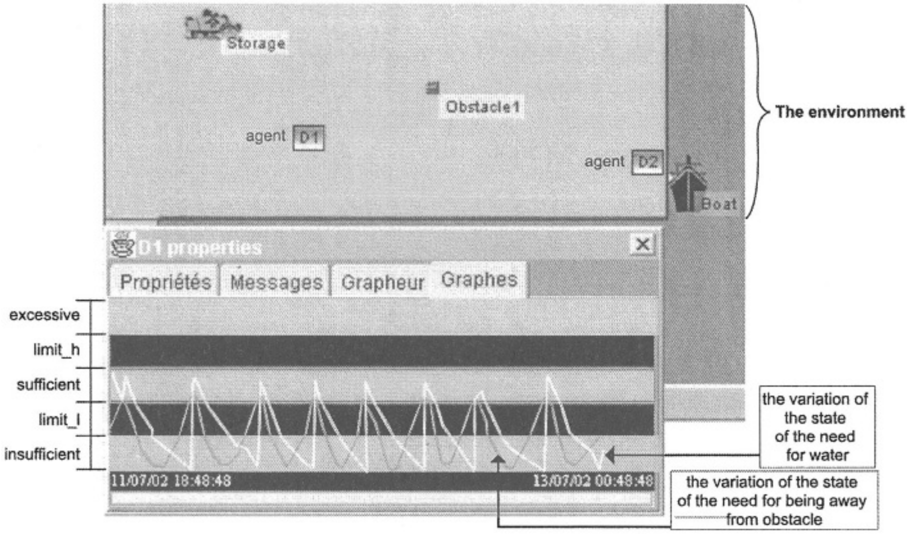


Fig. 6. A graphical result of a simulation. Here, we follow the need to drinking and the one to being away from obstacle

motivations) it is realized. We can see for example the case of joint intentions models [10] or some Belief-Desire-Intention (BDI) models [16]. The notion of *desire* in BDI is a first attempt at considering motivations. It is even called "motivational attitude" in [5]. However, this notion is only an abstract representation of natural motivations, at a higher level. The motivations themselves are not physically integrated into agents.

In some models about hybrid agents, works are not limited to only rational factors (planning, evaluation ...) but start to integrate the reaction to the dynamics of the environment. The most studied case is that of obstacle avoidance behavior which is primarily due to the natural preserving instinct. Among interesting examples, we may mention the works of [18] in which airplane pilots avoid each other, or even, that of [3] where agents try to reach their goals while avoiding fires (being burnt is related to a physiological factor). In these models, we can progressively see the manifestation of natural concepts. However, like in cognitive models with the notion of *desire,* motivations are also here considered according to the case-study only. These models do not contain a generic specification of the motivations.

This work is an attempt to overcome these situations.

Compared to [2], our more recent work, we have seen many evolutions in this paper:

- the list of criteria used for managing the need importance is added and some adjustments are fixed (see details in Section 6.2).
- the possible state of the need is also updated by adding the *excessive* state (Section 3.2). By doing so, our model is closer to the situation

– conceptually, our previous work does not greatly develop the discussion of the hybrid agent paradigm while in this paper, it does. It only outlined the hybridism notion and the work was rather focused on the need importance as well as the action selection mechanism.

9 Conclusion

9.1 Summary

We present in this paper a model of a hybrid agent based on basic animal/human needs whose satisfaction constitutes the basic motivations of agent behavior in individual and social environment. This research aims to explicitly represent humans and animals in a generic way their common integrated concepts: the natural needs. We adopt this approach unlike many current hybrid models that are focused on the study of agent behavior or agent plan in the composition of their hierarchical structure. Moreover, the pyramid we use for this work is a result of a real world investigation of the psychologist Abraham Maslow and as such, each level has a specific conceptual semantic not based on pure hypothesis. We treat both individual and social aspects of these needs.

9.2 Future Work

Despite this ongoing research, much functionality will still be investigated in the near future.

Modeling the cognitive part. Currently, the model is rather focused on the reactive than the cognitive part of the hybrid concept. Thus, the formalization of agent knowledge will be improved. In fact, in multi-agent paradigm, the information that an agent has about its environment is generally partial [6] unlike our assumption in this paper that an agent has knowledge of the (whole) need state of the pyramid of others. For instance, it may happen that for one or another reason, agents intentionally hide their real need states. Consequently, there is no real certitude in some information about them, there is just an assumption or a belief. In a general way, the concepts of BDI, i.e. *belief, desires* and *intention* (to hide some states) should be considered in future work. For this purpose, works such as [9,17] seem to be interesting references for us.

Furthermore, by studying this cognitive part, we can extend the criteria in the determination of the need importance discussed in Section 6. Indeed, we also consider criteria determined at a purely cognitive level, for instance resulting from an individual or collective organization, roles, negotiation, etc. between cognitive agents. Example about role: the leader of an organization may need more respect (esteem) than ordinary members. In a generally way, there will be particular rules which explicitly state that a given need PN_x must be considered first before a need PN_y.

Miscellaneous. Additional work will also be considered in the improvement of the model:

- integrating the general state of agents: we currently model the state of a given PN of an agent but not the general state of the pyramid, according to the set of all needs in it. This generalization is important because it allows the user to determine the global state of the agent itself
- improving the social modeling: in a real situation, the *help* action mentioned in Section 4.1 is not actually systematic. Generally, additional parameters must be taken into account before helping (the degree of relationship between the two dockers, individual objective of each docker, etc.). Moreover, such satisfaction of others' needs must also take into account the state of its own need before the decision.
- extending the notion of love as presented in Section 4.1: being loved by other agents such as family, friends, etc.

Acknowledgements. This research is funded by the "University of the Indian Ocean/Commission of the Indian Ocean/European Union" research program. We thank all people who are working there. We are also thankful to everybody who improved the quality of this paper, particularly the anonymous reviewers for their helpful suggestions and illuminating comments.

References

1. Andriamasinoro F., Courdier R. (2001). *Un Modèle Dynamique de Comportement Agents à base de Besoins.* In Amal El-Fallah Segrouchni, Laurent Magnin (eds.) : Journées Francophones sur l'Intelligence Artificielle Distribuée et les Systèmes Multi-Agents (JFIADSMA'01). Hermès Editions. November 12–15, Montreal, Canada.
2. Andriamasinoro F., Courdier R. (2002). *The Basic Instinct of Autonomous Cognitive Agents.* International Conference on Autonomous Intelligent System (ICAIS'2002), February **12th-15th**, Geelong, Australia
3. Au S., Liang J., Parameswaran N (1998). *Progressive plan execution in a dynamic world* in Dynamic and Uncertain Environments Workshop in Artificial Intelligent Planning, Eds. *Ralph Bergmann, Alexdrader Kott,* Pittsburgh USA, AAAI.
4. Bonabeau E., Theraulaz G. (1994). *L'Intelligence Collective.* Edition Hermès.
5. Brazier F., Dunin-Keplicz B., Treur J., Verbrugge R. (1999). *Beliefs, Intentions and DESIRE. Modeling internal Dynamic behavior of BDI agents.* In JJ Meyer, PY Schobbens (eds.) Formal Models of Agents: ESPRIT project Modelage final workshop, Lecture Notes in AI, volume 1760, Springer, pp 36–56.
6. Ferber J. (1997). *Les systèmes multi-agents, vers une intelligence collective.* Inter-Editions.
7. Ferguson, I. A. (1992). *TouringMachines: An Architecture for Dynamic, Rational, Mobile Agents.* PhD thesis, Clare Hall, University of Cambridge, UK.
8. Gershenson Carlos, González Pedro Pablo, Negrete Jose Martinez (2000). *Action Selection Properties in a Software Simulated Agent.,* in Cairó et. al. (Eds.) MICAI 2000: Advances in Artificial Intelligence. Lecture Notes in Artificial Intelligence 1793, pp. 634–648. Springer-Verlag.

9. Grosz B. J, Kraus S. (1996). *Collaborative plans for complex group actions* in Artificial Intelligence 86, pp.269–357.
10. Jennings N. R (1996) *Coordination Techniques for Distributed Artificial Intelligence,* in Foundations of Distributed Artificial Intelligence (eds. G. M. P. O'Hare and N. R. Jennings), Wiley.
11. Kurihara S., Aoyagi S., Onai R.(1997). *Adaptive Selection of Reactive/Deliberate Planning for the Dynamic Environment, - A Proposal and Evaluation of MRR-planning -.* In proceedings of the 8^{th} European Workshop on Modelling Autonomous Agents in a Multi-Agent World, (MAAMAW'97) Ronneby, Sweden.
12. Malec J. (2000) *On Augmenting Reactivity with Deliberation in a Controlled Manner.* Workshop on Balancing Reactivity and Social Deliberation in Multi-Agent Systems at the 14th European Conference on Artificial Intelligence (ECAI), Berlin, Germany.
13. Maslow A. H. (2000). *The Maslow business Reader.* Collins D. S. (eds.), J.Wiley & Sons, Inc.
14. Mavromichalis V. K., Vouros G. (2000). *ICAGENT: Balancing between Reactivity and Deliberation.* In Workshop on Balancing Reactivity and Social Deliberation in Multi-Agent Systems at the 14th European Conference on Artificial Intelligence (ECAI), Berlin, Germany.
15. Müller, J.P. (1994). *A conceptual model of agent interaction.* In Deen, S. M., editor, Draft proceedings of the Second International Working Conference on Cooperating Knowledge Based Systems (CKBS-94), pages 389–404, DAKE Centre, University of Keele, UK.
16. Panzarasa Pietro, Norman Timothy, Jennings Nicholas R. (1999) *Modeling sociality in the BDI framework* in Proceedings in 1^{st} Asia Pacific Conference on Intelligent Agent Technology, Hong-Kong.
17. Sullivan D., Grosz B., Kraus B. (2000): *Intention Reconciliation by Collaborative Agents.* 4^{th} International Conference on Multi-Agent Systems (ICMAS 2000), Boston USA, IEEE Computer Society Press.
18. Tambe M. (1996) *Executing Team Plans in Dynamic Multi-agent environments.* AAAI Fall Symposium on Plan Execution. Boston USA.
19. Wooldridge Michael (2002). *An Introduction To Multiagent System,* Published by John Wiley & Sons, March.

Author Index

This page intentionally left blank

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 2953: K. Konrad, Model Generation for Natural Language Interpretation and Analysis. XIII, 166 pages. 2004.

Vol. 2934: G. Lindemann, D. Moldt, M. Paolucci (Eds.), Regulated Agent-Based Social Systems. Proceedings, 2002. X, 301 pages. 2004.

Vol. 2930: F. Winkler, Automated Deduction in Geometry. VII, 231 pages. 2004.

Vol. 2923: V. Lifschitz, I. Niemelä (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 365 pages. 2004.

Vol. 2913: T.M. Pinkston, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2003. Proceedings, 2003. XX, 512 pages. 2003.

Vol. 2903: T.D. Gedeon, L.C.C. Fung (Eds.), AI 2003: Advances in Artificial Intelligence. Proceedings, 2003. XVI, 1075 pages. 2003.

Vol. 2902: F.M. Pires, S.P. Abreu (Eds.), Progress in Artificial Intelligence. Proceedings, 2003. XV, 504 pages. 2003.

Vol. 2892: F. Dau, The Logic System of Concept Graphs with Negation. XI, 213 pages. 2003.

Vol. 2891: J. Lee, M. Barley (Eds.), Intelligent Agents and Multi-Agent Systems. Proceedings, 2003. X, 215 pages. 2003.

Vol. 2882: D. Veit, Matchmaking in Electronic Markets. XV, 180 pages. 2003.

Vol. 2871: N. Zhong, **Z.W. Raś,** S. Tsumoto, E. Suzuki (Eds.), Foundations of Intelligent Systems. Proceedings, 2003. XV, 697 pages. 2003.

Vol. 2854: J. Hoffmann, Utilizing Problem Structure in Planing. XIII, 251 pages. 2003.

Vol. 2843: G. Grieser, Y. Tanaka, A. Yamamoto (Eds.), Discovery Science. Proceedings, 2003. XII, 504 pages. 2003.

Vol. 2842: R. Gavaldá, K.P. Jantke, E. Takimoto (Eds.), Algorithmic Learning Theory. Proceedings, 2003. XI, 313 pages. 2003.

Vol. 2838: **N. Lavrač,** D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Knowledge Discovery in Databases: PKDD 2003. Proceedings, 2003. XVI, 508 pages. 2003.

Vol. 2837: **N. Lavrač,** D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Machine Learning: ECML 2003. Proceedings, 2003. XVI, 504 pages. 2003.

Vol. 2835: T. Horváth, A. Yamamoto (Eds.), Inductive Logic Programming. Proceedings, 2003. X, 401 pages. 2003.

Vol. 2821: A. Günter, R. Kruse, B. Neumann (Eds.), KI 2003: Advances in Artificial Intelligence. Proceedings, 2003. XII, 662 pages. 2003.

Vol. 2807: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue. Proceedings, 2003. XIII, 426 pages. 2003.

Vol. 2801: W. Banzhaf, J. Ziegler, T. Christaller, P. Dittrich, J.T. Kim (Eds.), Advances in Artificial Life. Proceedings, 2003. XVI, 905 pages. 2003.

Vol. 2797: O.R. Zaïane, S.J. Simoff, C. Djeraba (Eds.), Mining Multimedia and Complex Data. XII, 281 pages. 2003.

Vol. 2792: T. Rist, R. Aylett, D. Ballin, J. Rickel (Eds.), Intelligent Virtual Agents. Proceedings, 2003. XV, 364 pages. 2003.

Vol. 2782: M. Klusch, A. Omicini, S. Ossowski, H. Laamanen (Eds.), Cooperative Information Agents VII. Proceedings, 2003. XI, 345 pages. 2003.

Vol. 2780: M. Dojat, E. Keravnou, P. Barahona (Eds.), Artificial Intelligence in Medicine. Proceedings, 2003. XIII, 388 pages. 2003.

Vol. 2777: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines. Proceedings, 2003. XIV, 746 pages. 2003.

Vol. 2752: G.A. Kaminka, P.U. Lima, R. Rojas (Eds.), RoboCup 2002: Robot Soccer World Cup VI. XVI, 498 pages. 2003.

Vol. 2741: F. Baader (Ed.), Automated Deduction – CADE-19. Proceedings, 2003. XII, 503 pages. 2003.

Vol. 2705: S. Renals, G. Grefenstette (Eds.), Text- and Speech-Triggered Information Access. VII, 197 pages. 2003.

Vol. 2703: O.R. Zaïane, J. Srivastava, M. Spiliopoulou, B. Masand (Eds.), WEBKDD 2002 - MiningWeb Data for Discovering Usage Patterns and Profiles. IX, 181 pages. 2003.

Vol. 2700: M.T. Pazienza (Ed.), Extraction in the Web Era. XIII, 163 pages. 2003.

Vol. 2699: M.G. Hinchey, J.L. Rash, W.F. Truszkowski, C.A. Rouff, D.F. Gordon-Spears (Eds.), Formal Approaches to Agent-Based Systems. IX, 297 pages. 2002.

Vol. 2691: **V. Mařík,** J.P. Müller, M. Pechoucek (Eds.), Multi-Agent Systems and Applications III. Proceedings, 2003. XIV, 660 pages. 2003.

Vol. 2684: M.V. Butz, O. Sigaud, P. Gérard (Eds.), Anticipatory Behavior in Adaptive Learning Systems. X, 303 pages. 2003.

Vol. 2671: Y. Xiang, B. Chaib-draa (Eds.), Advances in Artificial Intelligence. Proceedings, 2003. XIV, 642 pages. 2003.

Vol. 2663: E. Menasalvas, J. Segovia, P.S. Szczepaniak (Eds.), Advances in Web Intelligence. Proceedings, 2003. XII, 350 pages. 2003.

Vol. 2661: P.L. Lanzi, W. Stolzmann, S.W. Wilson (Eds.), Learning Classifier Systems. VII, 231 pages. 2003.

Vol. 2654: U. Schmid, Inductive Synthesis of Functional Programs. XXII, 398 pages. 2003.

Vol. 2650: M.-P. Huget (Ed.), Communications in Multiagent Systems. VIII, 323 pages. 2003.

Vol. 2645: M.A. Wimmer(Ed.), Knowledge Management in Electronic Government. Proceedings, 2003. XI, 320 pages. 2003.

Vol. 2639: G. Wang, Q. Liu, Y. Yao, A. Skowron (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Proceedings, 2003. XVII, 741 pages. 2003.

Vol. 2637: K.-Y. Whang, J. Jeon, K. Shim, J. Srivastava, Advances in Knowledge Discovery and Data Mining. Proceedings, 2003. XVIII, 610 pages. 2003.

Vol. 2636: E. Alonso, D. Kudenko, D. Kazakov (Eds.), Adaptive Agents and Multi-Agent Systems. XIV, 323 pages. 2003.

Vol. 2627: B. O'Sullivan (Ed.), Recent Advances in Constraints. X, 201 pages. 2003.

Vol. 2600: S. Mendelson, A.J. Smola (Eds.), Advanced Lectures on Machine Learning. IX, 259 pages. 2003.

Vol. 2592: R. Kowalczyk, J.P. Müller, H. Tianfield, R. Unland (Eds.), Agent Technologies, Infrastructures, Tools, and Applications for E-Services. XVII, 371 pages. 2003.

Vol. 2586: M. Klusch, S. Bergamaschi, P. Edwards, P. Petta (Eds.), Intelligent Information Agents. VI, 275 pages. 2003.

Vol. 2583: S. Matwin, C. Sammut (Eds.), Inductive Logic Programming. X, 351 pages. 2003.

Vol. 2581: J.S. Sichman, F. Bousquet, P. Davidsson (Eds.), Multi-Agent-Based Simulation. X, 195 pages, 2003.

Vol. 2577: P. Petta, R. Tolksdorf, F. Zambonelli (Eds.), Engineering Societies in the Agents World III. X, 285 pages. 2003.

Vol. 2569: D. Karagiannis, U. Reimer (Eds.), Practical Aspects of Knowledge Management. Proceedings, 2002. XIII, 648 pages. 2002.

Vol. 2560: S. Goronzy, Robust Adaptation to Non-Native Accents in Automatic Speech Recognition. XI, 144 pages. 2002.

Vol. 2557: B. McKay, J. Slaney (Eds.), AI 2002: Advances in Artificial Intelligence. Proceedings, 2002. XV, 730 pages. 2002.

Vol. 2554: M. Beetz, Plan-Based Control of Robotic Agents. XI, 191 pages. 2002.

Vol. 2543: O. Bartenstein, U. Geske, M. Hannebauer, O. Yoshie (Eds.), Web Knowledge Management and Decision Support. X, 307 pages. 2003.

Vol. 2541: T. Barkowsky, Mental Representation and Processing of Geographic Knowledge. X, 174 pages. 2002.

Vol. 2533: N. Cesa-Bianchi, M. Numao, R. Reischuk (Eds.), Algorithmic Learning Theory. Proceedings, 2002. XI, 415 pages, 2002.

Vol. 2531: J. Padget, O. Shehory, D. Parkes, N.M. Sadeh, W.E.Walsh (Eds.), Agent-Mediated Electronic Commerce IV. Designing Mechanisms and Systems. XVII, 341 pages. 2002.

Vol. 2527: F.J. Garijo, J.-C. Riquelme, M. Toro (Eds.), Advances in Artificial Intelligence - IBERAMIA 2002. Proceedings, 2002. XVIII, 955 pages. 2002.

Vol. 2522: T. Andreasen, A. Motro, H. Christiansen, H.L. Larsen (Eds.), Flexible Query Answering Systems. Proceedings, 2002. X, 383 pages. 2002.

Vol. 2514: M. Baaz, A. Voronkov (Eds.), Logic for Programming, Artificial Intelligence, and Reasoning. Proceedings, 2002. XIII, 465 pages. 2002.

Vol. 2507: G. Bittencourt, G.L. Ramalho (Eds.), Advances in Artificial Intelligence. Proceedings, 2002. XIII, 417 pages. 2002.

Vol. 2504: M.T. Escrig, F. Toledo, E. Golobardes (Eds.), Topics in Artificial Intelligence. Proceedings, 2002. XI, 427 pages. 2002.

Vol. 2499: S.D. Richardson (Ed.), Machine Translation: From Research to Real Users. Proceedings, 2002. XXI, 254 pages. 2002.

Vol. 2484: P. Adriaans, H. Fernau, M. van Zaanen (Eds.), Grammatical Inference: Algorithms and Applications Proceedings, 2002. IX, 315 pages. 2002.

Vol. 2479: M. Jarke, J. Koehler, G. Lakemeyer (Eds.), KI 2002: Advances in Artificial Intelligence. Proceedings, 2002. XIII, 327 pages. 2002.

Vol. 2475: J.J. Alpigini, J.F. Peters, A. Skowron, N. Zhong (Eds.), Rough Sets and Current Trends in Computing. Proceedings, 2002. XV, 640 pages. 2002.

Vol. 2473: A. Gómez-Pérez, V.R. Benjamins (Eds.), Knowledge Engineering and Knowledge Management Ontologies and the Semantic Web. Proceedings, 2002. XI, 402 pages. 2002.

Vol. 2466: M. Beetz, J. Hertzberg, M. Ghallab, M.E. Pollack (Eds.), Advances in Plan-Based Control of Robotic Agents. VIII, 291 pages. 2002.

Vol. 2464: M. O'Neill, R.F.E. Sutcliffe, C. Ryan, M. Eaton, N.J.L. Griffith (Eds.), Artificial Intelligence and Cognitive Science. Proceedings, 2002. XI, 247 pages. 2002.

Vol. 2448: P. Sojka, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue. Proceedings, 2002. XII, 481 pages 2002.

Vol. 2447: D.J. Hand, N.M. Adams, R.J. Bolton (Eds.), Pattern Detection and Discovery. Proceedings, 2002. XII, 227 pages. 2002.

Vol. 2446: M. Klusch, S. Ossowski, O. Shehory (Eds.), Cooperative Information Agents VI. Proceedings, 2002. XI, 321 pages. 2002.

Vol. 2445: C. Anagnostopoulou, M. Ferrand, A. Smaill (Eds.), Music and Artificial Intelligence. Proceedings, 2002. VIII, 207 pages. 2002.

Vol. 2443: D. Scott (Ed.), Artificial Intelligence: Methodology, Systems, and Applications. Proceedings, 2002. X, 279 pages. 2002.

Vol. 2432: R. Bergmann, Experience Management. XXI, 393 pages. 2002.

Vol. 2431: T. Elomaa, H. Mannila, H. Toivonen (Eds.), Principles of Data Mining and Knowledge Discovery. Proceedings, 2002. XIV, 514 pages. 2002.

Vol. 2430: T. Elomaa, H. Mannila, H. Toivonen (Eds.), Machine Learning: ECML 2002. Proceedings, 2002. XIII, 532 pages. 2002.

Vol. 2427: M. Hannebauer, Autonomous Dynamic Reconfiguration in Multi-Agent Systems. XXI, 284 pages. 2002.