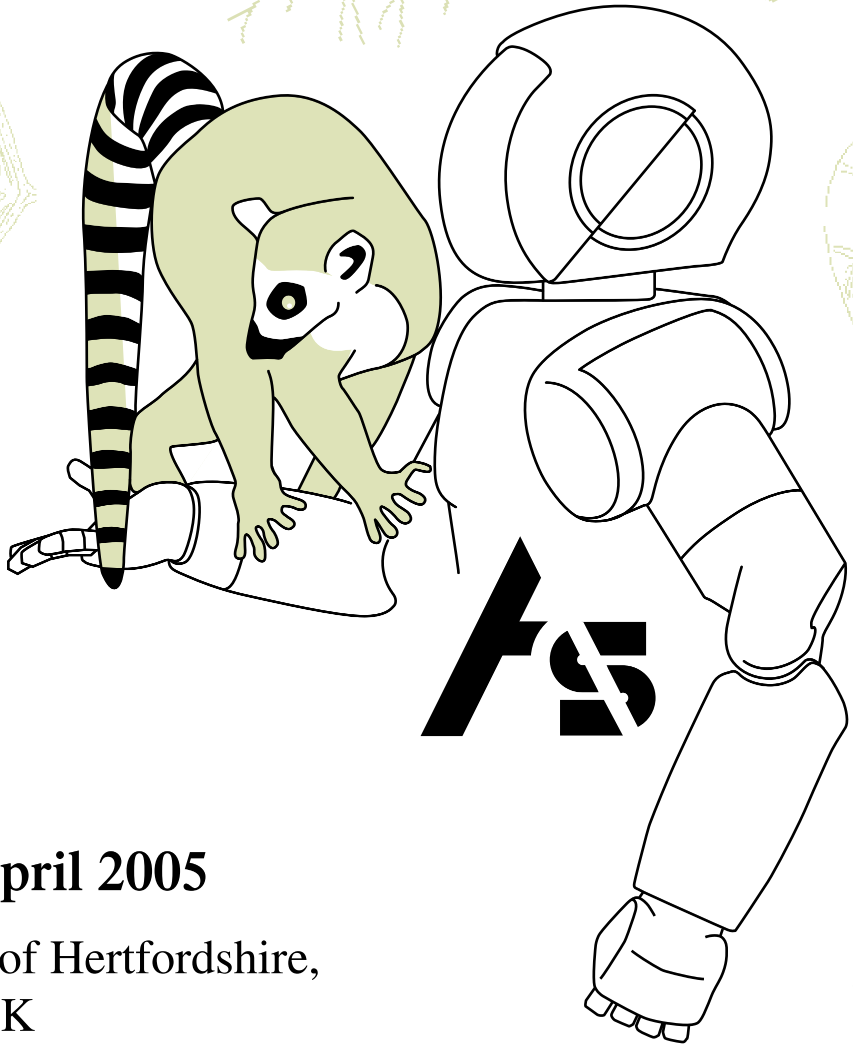
AISB'05: Social Intelligence and Interaction in Animals, Robots and Agents

Proceedings of the Symposium on Normative Multi-Agent Systems



12 - 15 April 2005

University of Hertfordshire,
Hatfield, UK

SSAISB 2005 Convention

AISB UH EPSRC
Engineering and Physical Sciences
Research Council

AISB'05 Convention

Social Intelligence and Interaction in Animals, Robots and Agents

12-15 April 2005
University of Hertfordshire, Hatfield, UK

Proceedings of the Symposium on

Normative Multi-Agent Systems

Published by

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour
www.aisb.org.uk

Printed by

The University of Hertfordshire, Hatfield, AL10 9AB UK
www.herts.ac.uk

Cover Design by Sue Attwood

ISBN 1 902956 47 6

AISB’05 Hosted by

The Adaptive Systems Research Group
adapsys.feis.herts.ac.uk

The AISB'05 Convention is partially supported by:

EPSRC Engineering and Physical Sciences
Research Council

The proceedings of the ten symposia in the AISB’05 Convention are available from SSAISB:

Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
1 902956 40 9

Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
1 902956 41 7

Third International Symposium on Imitation in Animals and Artifacts
1 902956 42 5

Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
1 902956 43 3

Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
1 902956 44 1

Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
1 902956 45 X

Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
1 902956 46 8

Normative Multi-Agent Systems
1 902956 47 6

Socially Inspired Computing Joint Symposium (Memetic theory in artificial systems & societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
1 902956 48 4

Virtual Social Agents Joint Symposium (Social presence cues for virtual humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)
1 902956 49 2

Table of Contents

The AISB'05 Convention - Social Intelligence and Interaction in Animals, Robots and Agents……… i
K.Dautenhahn

Symposium Preface - Normative Multi-Agent Systems……………………..……………………… iv
Guido Boella, Leon van der Torre, and Harko Verhagen

Introduction to Normative Multiagent Systems………………...…………………………………… 1
Guido Boella, Leender van der Torre, and Harko Verhagen

A Framework for the Design of Self-Regulation of Open Agent-based Electronic Marketplaces…….. 8
Christian S. Hahn, Bettina Fley, and Michael Florian

A Normative Framework for Agent-Based Systems………………………………………………….. 24
Fabiola López y Lopéz, Michael Luck, and Mark d'Inverno

Beyond BDI? Brandomian commitments for multi-agent communication…………………………… 36
Rodger Kibble

Formalizing Coalition Structures via Obligations…………………………………………………….. 48
Luigi Sauro

Impact of Multiple Normative Systems on Organization Performance of
International Joint Ventures……………………………………………………………………… 54
Tamaki Horii, Yan Jin, and Raymond E. Levitt

Increasing Software Infrastructure Dependability through a Law Enforcement Approach…………... 65
Gustavo Carvalho, Rodrigo Paes, Ricardo Choren, Paulo Alencar, and Carlos Lucena

My Agents Love to Conform: Emotions, Norms, and Social Control in Natural and
Artificial Societies…………………………………………………………………………………... 73
Christian von Scheve, Daniel Moldt, Julia Fix, and Rolf von Lüde

Normative KGP Agents: A Preliminary Report………………...………………………………….. 85
Fariba Sadri, Francesca Toni, and Kostas Stathis

On the Potential of Norm-Governed Behavior in Different Categories of Artificial Societies……….. 97
Paul Davidsson and Stefan Johansson

Ontological Aspects of the Implementation of Norms in Agent-Based Electronic Institutions………104
Davide Grossi, Huib Aldewereld, Javier Vázquez-Salceda, and Frank Dignum

Self-Organized Criticality and Norm Avalanches…………………………………………………….117
Matthew J. Hoffmann

Mapping of Deontic Operators to Abductive Expectations…………………………………………...126
*Marco Alberti, Marco Gavanelli, Evelina Lamma, Paola Mello, Paolo Torroni,
and Giovanni Sartor*

Towards Norm-Governed Self-Organising Networks………………………………………………...137
Lloyd Kamara, Jeremy Pitt, and Marek Sergot

The AISB’05 Convention
Social Intelligence and Interaction in Animals, Robots and Agents

Above all, the human animal is social. For an artificially intelligent system, how could it be otherwise?

We stated in our Call for Participation “The AISB’05 convention with the theme *Social Intelligence and Interaction in Animals, Robots and Agents* aims to facilitate the synthesis of new ideas, encourage new insights as well as novel applications, mediate new collaborations, and provide a context for lively and stimulating discussions in this exciting, truly interdisciplinary, and quickly growing research area that touches upon many deep issues regarding the nature of intelligence in human and other animals, and its potential application to robots and other artefacts”.

Why is the theme of Social Intelligence and Interaction interesting to an Artificial Intelligence and Robotics community? We know that intelligence in humans and other animals has many facets and is expressed in a variety of ways in how the individual in its lifetime - or a population on an evolutionary timescale - deals with, adapts to, and co-evolves with the environment. Traditionally, social or emotional intelligence have been considered different from a more problem-solving, often called "rational", oriented view of human intelligence. However, more and more evidence from a variety of different research fields highlights the important role of social, emotional intelligence and interaction across all facets of intelligence in humans.

The Convention theme *Social Intelligence and Interaction in Animals, Robots and Agents* reflects a current trend towards increasingly interdisciplinary approaches that are pushing the boundaries of traditional science and are necessary in order to answer deep questions regarding the social nature of intelligence in humans and other animals, as well as to address the challenge of synthesizing computational agents or robotic artifacts that show aspects of biological social intelligence. Exciting new developments are emerging from collaborations among computer scientists, roboticists, psychologists, sociologists, cognitive scientists, primatologists, ethologists and researchers from other disciplines, e.g. leading to increasingly sophisticated simulation models of socially intelligent agents, or to a new generation of robots that are able to learn from and socially interact with each other or with people. Such interdisciplinary work advances our understanding of social intelligence in nature, and leads to new theories, models, architectures and designs in the domain of Artificial Intelligence and other sciences of the artificial.

New advancements in computer and robotic technology facilitate the emergence of multi-modal "natural" interfaces between computers or robots and people, including embodied conversational agents or robotic pets/assistants/companions that we are increasingly sharing our home and work space with. People tend to create certain relationships with such socially intelligent artifacts, and are even willing to accept them as helpers in healthcare, therapy or rehabilitation. Thus, socially intelligent artifacts are becoming part of our lives, including many desirable as well as possibly undesirable effects, and Artificial Intelligence and Cognitive Science research can play an important role in addressing many of the huge scientific challenges involved. Keeping an open mind towards other disciplines, embracing work from a variety of disciplines studying humans as well as non-human animals, might help us to create artifacts that might not only do their job, but that do their job right.

Thus, the convention hopes to provide a home for state-of-the-art research as well as a discussion forum for innovative ideas and approaches, pushing the frontiers of what is possible and/or desirable in this exciting, growing area.

The feedback to the initial Call for Symposia Proposals was overwhelming. Ten symposia were accepted (ranging from one-day to three-day events), organized by UK, European as well as international experts in the field of Social Intelligence and Interaction.

- Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
- Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
- Third International Symposium on Imitation in Animals and Artifacts
- Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
- Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
- Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
- Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
- Normative Multi-Agent Systems
- Socially Inspired Computing Joint Symposium (consisting of three themes: Memetic Theory in Artificial Systems & Societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
- Virtual Social Agents Joint Symposium (consisting of three themes: Social Presence Cues for Virtual Humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)

I would like to thank the symposium organizers for their efforts in helping to put together an excellent scientific programme.

In order to complement the programme, five speakers known for pioneering work relevant to the convention theme accepted invitations to present plenary lectures at the convention: Prof. Nigel Gilbert (University of Surrey, UK), Prof. Hiroshi Ishiguro (Osaka University, Japan), Dr. Alison Jolly (University of Sussex, UK), Prof. Luc Steels (VUB, Belgium and Sony, France), and Prof. Jacqueline Nadel (National Centre of Scientific Research, France).

A number of people and groups helped to make this convention possible. First, I would like to thank SSAISB for the opportunity to host the convention under the special theme of *Social Intelligence and Interaction in Animals, Robots and Agents*. The AISB'05 convention is supported in part by a UK EPSRC grant to Prof. Kerstin Dautenhahn and Prof. C. L. Nehaniv. Further support was provided by Prof. Jill Hewitt and the School of Computer Science, as well as the Adaptive Systems Research Group at University of Hertfordshire. I would like to thank the Convention's Vice Chair Prof. Chrystopher L. Nehaniv for his invaluable continuous support during the planning and organization of the convention. Many thanks to the local organizing committee including Dr. René te Boekhorst, Dr. Lola Cañamero and Dr. Daniel Polani. I would like to single out two people who took over major roles in the local organization: Firstly, Johanna Hunt, Research Assistant in the School of Computer Science, who efficiently dealt primarily with the registration process, the AISB'05 website, and the coordination of ten proceedings. The number of convention registrants as well as different symposia by far exceeded our expectations and made this a major effort. Secondly, Bob Guscott, Research Administrator in the Adaptive Systems Research Group, competently and with great enthusiasm dealt with arrangements ranging from room bookings, catering, the organization of the banquet, and many other important elements in the convention. Thanks to Sue Attwood for the beautiful frontcover design. Also, a number of student helpers supported the convention. A great team made this convention possible!

I wish all participants of the AISB’05 convention an enjoyable and very productive time. On returning home, I hope you will take with you some new ideas or inspirations regarding our common goal of understanding social intelligence, and synthesizing artificially intelligent robots and agents. Progress in the field depends on scientific exchange, dialogue and critical evaluations by our peers and the research community, including senior members as well as students who bring in fresh viewpoints. For social animals such as humans, the construction of scientific knowledge can't be otherwise.

Dedication:

I am very confident that the future will bring us increasingly many instances of socially intelligent agents. I am similarly confident that we will see more and more socially intelligent robots sharing our lives. However, I would like to dedicate this convention to those people who fight for the survival of socially intelligent animals and their fellow creatures. What would 'life as it could be' be without 'life as we know it'?

Beppu, Japan.

Kerstin Dautenhahn

Professor of Artificial Intelligence,
General Chair, AISB'05 Convention *Social Intelligence and Interaction in Animals, Robots and Agents*

University of Hertfordshire
College Lane
Hatfield, Herts, AL10 9AB
United Kingdom

Symposium Preface

Normative Multi-Agent Systems

1st International Symposium on Normative Multiagent Systems (NorMAS2005)

SYMPOSIUM OVERVIEW

NorMAS2005 is a two day symposium part of the 2005 AISB convention. The general theme for the AISB 2005 convention is "Social Intelligence and Interaction in Animals, Robots and Agents". It is held from April 12 to April 2005 at the University of Hertfordshire, Hatfield, England. AISB conventions are organized by the largest AI society in the United Kingdom, SSAISB which stands for Society of the Study of Artificial Intelligence and the Simulation of Behaviour. NorMAS2005 will take place on Tuesday, April 12th and Wednesday, April 13th.

The best papers of the symposium will be selected for publication in special issues of Computational Intelligence and Computational & Mathematical Organization Theory.

DESCRIPTION OF NorMAS

Norms are essential for artificial agents that are to display behaviour comparable to human intelligent behaviour or collaborate with humans, because the use of norms is the key of human social intelligence. Norms play a central role in many social phenomena such as coordination, cooperation, decision-making, etc. There is an increasing interest in the role of norms in societies, both inside as outside the agent community. Now the time is ripe for a symposium focussing on this central sociological concept given that the field of (multi)agent research is moving more and more from the individual, cognitive focussed agent models to models of socially situated agents. NorMAS therefore focuses on normative multiagent systems.

Normative multiagent systems combine theories and frameworks for normative systems with multiagent systems. Thus, these systems provide a promising model for human and artificial agent coordination, because they integrate norms and individual intelligence. They are a prime example of the use of sociological theories in multiagent systems, and therefore of the relation between agent theory and the social sciences, e.g., sociology, philosophy, economics, legal science, etc.

NorMAS2005, as part of AISB2005, will provide an excellent opportunity to meet researchers studying norms in cognitive science, social sciences, agent theory, computer science, philosophy, etc. to discuss the current state and identify potential future directions and research issues.

TOPICS OF INTEREST

The topics of this symposium include, but are not restricted to, the following issues:

- multiagent or society level:
    - balancing dynamics and statics at the agent (micro) and agent society (macro) level
    - coordination based on normative multiagent systems
    - emergence of conventions, norms, roles, and normative multiagent systems
    - combining conventions with regulative, permissive, constitutive and other types of norms
    - relation between NorMAS and contracts, security, and (electronic) institutions
- agent level:
    - alternatives to and extensions of the homo economicus and BDI logics
    - extending logical frameworks to encompass norms in agent decision making
    - how to implement theories of norms in artificial agents
- applications of NorMAS:
    - multiagent social simulation models containing norms
    - mixing artificial and human agents in hybrid social systems

PROGRAMME COMMITTEE

Guido Boella - Dipartimento di Informatica, Universita' di Torino (co-chair)
Cristiano Castelfranchi - Institute of Cognitive Sciences and Technologies (ISTC), Italy
Paul Davidsson - BTH, Sweden
André Meyer - TNO, Netherlands
Maria Fasli - Essex University, UK
Leendert van der Torre - CWI Amsterdam, Netherlands (co-chair)
Harko Verhagen - DSV, KTH/SU, Sweden (co-chair)

SYMPOSIUM WEBSITE

http://normas.di.unito.it/zope/aisb05/

AISB 2005 CONVENTION WEBSITE

http://aisb2005.feis.herts.ac.uk/

Introduction to Normative Multiagent Systems

Guido Boella[⋆]
[⋆]Dipartimento di Informatica
Università di Torino
Italy
guido@di.unito.it

Leendert van der Torre[†]
[†]CWI Amsterdam and
Delft University of Technology
The Netherlands
torre@cwi.nl

Harko Verhagen[‡]
[‡]Dept of Computer and Systems Sciences
Stockholm University \ KTH,
Forum 100, SE-16440 Kista, Sweden
verhagen@dsv.su.se

Abstract

In this paper we give a short introduction to the emerging area of normative multiagent systems by presenting definitions and examples.

1 Introduction

Normative multiagent systems as a research area is best defined as the intersection of two established fields: normative systems and multiagent systems. This still leaves a lot of room for interpretation, as there are various definitions of both areas. However, it is clear that the intersection involves several issues which are highly relevant nowadays, such as coordination, multiagent organizations and agent security. This is witnessed both in the field of normative systems, where the last workshop on deontic logic in computer science (ΔEON04) had as topic "application to multiagent systems", and in the field of multiagent systems, where various subfields now are addressing normative systems, such as multiagent organization or regulated societies.

In this paper we make some observations with respect to the emerging research area of normative multiagent systems by addressing the following questions:

1. What is a normative multiagent system?

2. What are prototypical examples of normative multiagent systems?

To answer questions, we consider normative multiagent systems at the border of agent theory - both multiagent systems and autonomous agents - and the social sciences - sociology, philosophy, economics, et cetera. The forces at this border are considered in two directions: how do the social sciences influence agent theory, and how does agent theory influence the social sciences?

Social sciences ⇒ Agent theory. The social sciences are regularly used in the development of theories and models of multiagent systems. It is used in two ways. The first and most obvious way is the use in agent theory of concepts developed in the social sciences, such as co-ordination, organization, convention, norm, trust, et cetera. A second and less popular way, but often at least as useful, is to contrast agent theory with social theory, based on the distinctions between artificial systems and humans. For example, humans cannot be programmed such that they never violate a norm or always co-operate, but artificial systems can.

Agent Theory ⇒ Social sciences. According to Castelfranchi (1998), agent theory should also produce theories, models, and experimental, conceptual and theoretical new instruments, which can be used to revise and develop the social sciences. He summarises this point by stating that agent theory - and the related area of artificial intelligence - is not just an engineering discipline, but it is also a science.

The layout of this paper follows the questions. First we give some definitions, and then we discuss some examples.

2 Definitions

In this section we first introduce normative systems, then we consider norms in sociology, and finally we consider multiagent systems.

2.1 Normative systems

Normative systems have traditionally been studied by legal philosophers like Alchourròn and Bulygin (1971).

Meyer and Wieringa, who founded in 1991 the deontic logic in computer science workshops (known as the ΔEON workshops), define normative systems as

> "systems in the behavior of which norms play a role and which need normative concepts in order to be described or specified" (Meyer and Wieringa, 1993, preface).

They also explain why normative systems are intimately related with deontic logic:

> "Until recently in specifications of systems in computational environments the distinction between normative behavior (as it *should be*) and actual behavior (as it *is*) has been disregarded: mostly it is not possible to specify that some system behavior is non-normative (illegal) but nevertheless possible. Often illegal behavior is just ruled out by specification, although it is very important to be able to specify what should happen if such illegal but possible behaviors occurs! Deontic logic provides a means to do just this by using special modal operators that indicate the status of behavior: that is whether it is legal (normative) or not."

Deontic logic was founded by von Wright in 1951 as the formal study of ought. His main idea was that the deontic modalities of obligation and permission were related to each other in the same way as the alethic ones of necessity and possibility are related to each other. Thus, the logic of "It is obligatory to see to it that x" together with "it is permitted to see to it that x" is analogous to the logic of "it is necessary that x" together with "it is possible that x". Permissions are defined in terms of obligations just like possibility it defined in terms of necessity (it is not necessary that the absence of x is the case). However, whereas for necessity often a universal relation is taken (KD45), von Wright gave deontic logic a much weaker system (a weakened version of KD).

Another relation between deontic logic and alethic modal logic was pioneered by Anderson, who defined deontic logic in alethic modal logic with an additional constant (known as Anderson's reduction): "the intimate connection between obligations and sanctions in normative systems suggests that we might profitably begin by considering some penalty or sanction S, and define obligations as: p is obligatory if its falsity entails the sanction S". This was formalized by (a more complex variant of) $O(p) = \Box(\neg p \rightarrow S)$. When people objected that not all violations are sanctioned, Anderson replied that "S just means something bad has happened or a violation has occurred". Much later, Meyer (1988) used a similar reduction to dynamic logic.

Unfortunately, it soon became apparent that it was unclear how to formalize conditionals or rules in these modal systems, and many examples formalized in deontic logic had counterintuitive conclusions, known as the deontic paradoxes. The most notorious ones of them, the so-called contrary-to-duty paradoxes, are concerned with revision of obligations in case of violations.

Von Wright intended that the propositions of his deontic logic referred to actions, it was a logic that described what an actor has to do. Most people reinterpreted it as a system without any agency, but in seventies and eighties many temporal and action logics were introduced. Jones and Carmo (2001) recently define normative systems as, what we will call here, normative multiagent systems:

> "Sets of agents whose interactions are norm-governed; the norms prescribe how the agents ideally should and should not behave. [...] Importantly, the norms allow for the possibility that actual behavior may at times deviate from the ideal, i.e., that violations of obligations, or of agents' rights, may occur."

This does not mean, however, that this agrees with views on norms for multiagent systems. The most common view on norms in multiagent systems is that norms are constraints on behavior via social laws; an alternative view studies cognitively grounded norms. Conte and Castelfranchi (1995) mention three kinds of norms: norms as constraints on behavior, norms as ends (or goals) and norms as obligations. In the following section, we consider social systems.

2.2 Norms in sociology

In this section we analyze the use of norms within sociology. In sociology, the use of norms has been out of fashion since the 1950's, apart from game theoretically inspired research. In a recent article, Therborn (2002) gives some reasons for this while at the same time presenting an overview of the use of and ideas about norms within sociology during the 1900's.

Within sociology, one can distinguish several action models. Here we present the overview developed by Habermas (1984) of the sociological action models.

The first action model is teleological action. Agents are goal directed and try to maximize their choice of means to obtain a goal. This is the rational choice model. The central issue in this action model is the choices the agent makes between different action alternatives, based on maximizing utility. Agents can thus - try to influence the world, and the rationality of the behavior of the agents can be evaluated with respect to the efficiency of their behavior. Adding other agents, with respect to whom the agent acts in a strategic manner (strategical action model), to the decision making model does not change the ontological principles. The agents may need to model the desires and actions of the other agents but these are still part of the objective world of existing states of affairs. Agents act with respect to this world according to their beliefs about the existing states of affairs and their intentions to bring about desired states of affairs in that world.

The second action model is the normatively regulated action model. Social agents are assumed to belong to a group and follow the norms that are obliged to be followed by members of that group. Following norms is taken as to behave according to expectations. The objective reality is extended by a social reality of obliging norms (acknowledged as such by the group). The rationality of the behavior of an agent is not only related to the objective reality (teleological and strategical action model), but also to the social reality. The conformity between the norms of the group and the behavior of the agents and the relation between the norms and the generalized interests of the agents (and thus if it is wise of the agents to confirm to those norms) are part of this social rationality. Agents act with respect to an objective world and a social world, namely the normative context that defines the possible interactions and legitimate interagent relationships between the agents.

The third action model is the dramaturgical action model. In this action model the inner world of the agents is considered. Based on the dramaturgical analysis of social life as developed by Goffman (1959), this action model has as a core the presentation of the self of an agent to an audience. This representation of the self may or may not be truthful. The agent makes use of the fact that its inner self is only admissible to itself. The inner self is defined as the constellation of beliefs, desires, intentions, feelings, and needs of an agent. Habermas views this inner self as a reality in its own right. When presented in a truthful and authentic way, and at the same time connected to the shared evaluation criteria and interpretations of needs, the subjective point of view of the agent can gain an intersubjective value. Truthful is not the same as true in objective sense, opening the door for lying and manipulation or *insincerity*. Agents act with respect to an objective world and a subjective world formed by the totality of subjective experience to which the agent has a privileged access. Examples of application of this action model in the field of MAS include lying agents and believable agents.

The fourth and final action model is the communicative action model. This action model unites the three functions of language specified by the three previous action models. In the strategical action model, language is used by an agent to reach its own goals possibly via influencing other agents by use of language, the normative action model uses language to actualize already existing normative agreements and the dramaturgical model uses language to allow for one to express oneself. In the communicative action model, language is used to bring about mutual understanding on all three previous levels. The agents use language to claim the truth of their utterances, the normative correctness of their speech acts in view of the context, and the sincerity of their intentions being formulated. Testing for rationality of actions is here no longer the privilege of the observer, but is done by the agents themselves to realize a common definition of the situation described in terms of relations between the speech act and the three worlds (i.e., the objective, social, and subjective world) this speech act has relations with. In the cooperative process of interpretation, all participating agents have to incorporate their own interpretation with that of the other agents so that the agents have a sufficiently shared view of the external (i.e., objective and social) world in order to coordinate their actions while pursuing their own goals.

For this article, we will focus upon the normative action model. Following Therborn Therborn (2002) we make some helpfull distinctions. For one, it cannot be stated that all actions that comply with norms can be called normative action in a more strict sense. Different reasons for complying with norms exist.

Normfollowing can be instrumental, where the reasons for compying are either for direct rewards or to avoid the costs of violation. Other reasons can be more socially oriented, such as the desire to belong to a group, not to loose face or esteem, avoid legal punishment, etc. i.e., socially instrumental reasons. Normative action is action where norms are followed for their own sake. This may be out of habit, in an unconscious way, or in a conscious or rational way, based upon an analysis of the consequences of actions within the social world. Other reasons for normfollowing include identification (e.g., with a group, institution or nation) and normfollowing out of self-respect. These reasons represent different levels of internatilization of norms. Norms correlated with self-respect are deeply rooted within the personality of the agent, whereas the identification norms are more shallowly rooted.

We may also look at norms from a functional point of view, what do norms result in? For one we have norms that are of a constitutive nature, they define the agent's membership in a system of action, ahd the system of action at large. Another funcion of norms is regulation, describing what members of a social system must and must not do. Thirdly, norms may have a distributive function, that is how rewards, costs and risks are to be divided among the social system's members.

Independent of the various types of norms, some main issues involved with discussions of norms are norm conformity and norm violoation and the dynamics of norms. If agents are to comply with norms, Norm conformity and violation issues

One characteristic that MAS research and social science, and sociology in particular, share is the interest in the relation between micro-level behaviour and macro-level effects. In MAS research, this boils down to the question “How to ensure efficiency at the level of the multiagent system whilst respecting individual autonomy?”. Possible solutions to this problem comprise of:

- use of central control
- internalized control, e.g. the use of social laws Shoham and Tennenholtz (1992).
- structural coordination as proposed in Ossowski (1999)
- a set of norms and learning at all levels, including the level of norms, based on reflecting upon the results of actions.

2.3 Multiagent systems

In agent research or agent system development, omnipotent agents hardly exist, in fact if an agent can be omnipotent, we can do without the concept of agents. Agents are used in software development since knowledge is limited and local autonomy is needed. Also, the single agent paradigm is not really an implementation of agents. The basic agent definition as sketched out by ,e.g., Wooldridge (2002) states that an agent has the following characteristics:

> ” it is a computer system that is situated in some environment and that is capable of autonomous action in this environment in order to meet its design objectives ”

where autonomy means control over behaviour and internal state. This definition is further developed defining of weak versus strong agency:

> ” Intelligent agent (Wooldridge 2002) - weak agency: an intelligent agent is capable of flexible autonomous action
>
> - flexibility meaning: reactivity: interact with environment pro-activeness: take initiative
> - social ability: interact with other agents/ co-operation o autonomy meaning: operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state ”

Strong agency uses anthropomorphic notions based on mentalistic properties such as beliefs, desires, intentions, rationality and emotions.

2.4 Norms and multiagent systems

Norms are essential for artificial agents that are to display behavior comparable to human intelligent behavior or collaborate with humans, because the use of norms is the key of human social intelligence. Norms play a central role in many social phenomena such as coordination, cooperation, decision-making, etc. There is an increasing interest in the role of norms in societies, both inside as outside the agent community. The field of (multi)agent research is moving more and more from the individual, cognitive focussed agent models to models of socially situated agents.

Normative multiagent systems combine theories and frameworks for normative systems with multiagent systems. Thus, these systems provide a promising model for human and artificial agent coordination, because they integrate norms and individual intelligence. They are a prime example of the use of sociological theories in multiagent systems, and therefore of the relation between agent theory and the social sciences, e.g., sociology, philosophy, economics, legal science, etc.

Below we describe our work on norm autonomous agents and our work on normative multiagent systems.

2.4.1 Norm autonomous agents

In the framework developed in Verhagen (2000), norm autonomous agents are described. In short, these agents are based upon cognitive (or goal autonomous agents as developed by Conte & Castelfranchi Conte and Castelfranchi (1995)) and are extended with norms. The agents are part of a normative framework, and at the same time reason about and are able to influence these norms. In this sense norm autonomous agents span the institutional (or even inter-institutional level) where norms get their meaning, the inter-individual level (groups of where norms are produced), and the individual level (where the individual decision making is taking place). These agents choose which goals are legitimate to pursue, based on a given system of norms. The agent has the autonomy of generating its own goals and to choose which it is going to pursue. Besides, the agent is equipped to judge the legitimacy of its own goals and other agents' goals. When a goal conflict arises (not to be confused with interest conflict), the agent may change its norm system thereby changing priorities of goals, abandoning a goal, changing a goal, generating another goal, etc. The reasoning capability of these agents at the level of norms is called normative reasoning. Norm autonomous agents generate norms they can use to evaluate states of the world in terms of whether or not they could be legitimate interests. Legitimacy is a social notion and is in the end determined by the norms of the agent with respect to the agent society it is part of.

2.4.2 Normative multiagent systems

Castelfranchi (1998, 2000) defines several social viewpoints on multiagent systems, in which he conceptualizes a multiagent system in terms of respectively the mind of the agent, the power of the agent, the dependencies between agents, and groups or coalitions. Moreover, he defines abstraction relations between them, related to emergence of social structures from individual agents.

In the Boella and van der Torre (to appear) model of normative multiagent systems, part of Castelfranchi's model is formalized in terms of rule based systems, based on deontic logic of van der Torre (2003) and Broersen et al. (2002, to appear)'s BOID architecture. These models are used to model a variety of phenomena, such as virtual communities, co-operation in groups, contracts, and constitutive norms.

The formal characteristic of the model is that it combines a logical framework with decision-theoretic and game-theoretic mechanisms, such that the behavior of the agents as well as the system can be formalized. For example, in the model of co-operation within groups the game-theoretic concepts are used to model the property of agents in a group are committed to mutual responsiveness, that is, they monitor the behavior of other agents and help them if possible to reach their goals.

The development of the model is driven by examples found in the social and legal literature. An example is Beccaria's argument that high penalties for minor offences increases the total set of norm violations, because once an agent has committed a violation he or she is no longer constrained to commit more violations. The model combines a logical framework to represent and reason about norms and norm violations, with a decision-theoretic mechanism to explain the behavior of the violator.

3 Examples

3.1 Coordination and cooperation

Shoham and Tennenholtz (1992) introduce artificial social systems to coordinate multiagent systems, using a kind of norms called *social laws*.

> In multiagent systems be they human societies or distributed computing systems different agents, people or processes, aim to achieve different goals and yet these agents must interact either directly by sharing information and services or indirectly by sharing system resources. In such distributed systems it is crucial that the agents agree on certain rules in order to decrease conflicts among them and promote cooperative behavior. Without such rules even the simplest goals might become unattainable by any of the agents or at least not efficiently attainable. Just imagine driving in the absence of traffic rules. These rules

> strike a balance between allowing agents sufficient freedom to achieve their goals and restricting them so that they do not interfere too much with one another.

They consider the possibility of limiting the agents to a subset of the original strategies of a given game thus inducing a subgame of the original one. They call such a restriction a social constraint if the restriction leaves only one strategy to each agent. Some social constraints are consistent with the principle of individual rationality in the sense that it is rational for agents to accept those assuming all others do as well.

A discussion in artificial social systems is whether social laws are hard or soft constraints. The distinction between hard and soft constraints corresponds to the distinction between preventative and detective control systems. In the former a system is built such that violations are impossible (you cannot enter metro station without a ticket) or that violations can be detected (you can enter train without a ticket but you may be checked and sanctioned).

3.2 Multiagent organizations

Organizations embody a powerful way to coordinate complex behavior in human society. Different models of organizations exist, from bureaucratic systems based on norms to competitive systems based on markets. Moreover, organizational concepts allow to structure the behavior of complex entities in a hierarchy of encapsulated entities: departments structured in roles, organizations structured in departments, and inter-organizational coordination structured in organizations. Organizations specify the interaction and communication possibilities of each of these entities, abstracting from the implementation of their behavior. Since these entities are autonomous, they can only be coordinated exogenously.

Organizational models have been popular in the last years in agent theory for modeling coordination in open systems, where departments and organizations are modeled as autonomous entities. This is also due to the need to ensure social order within MAS applications like Web Services, Grid Computing, and Ubiquitous Computing. In these settings, openness, heterogeneity, and scalability pose new challenges on traditional MAS organizational models. It becomes necessary to integrate organizational and individual perspectives and to promote the dynamic adaptation of models to organizational and environmental changes. Nowadays, practical applications of agents to organizational modeling are being widely developed.

Moreover, organizational concepts are used frequently for coordination purposes in different areas of Computer Science. For example, roles are used in access control, conceptual modeling, programming languages and patterns. Contracts are used in design by contract, and services are used in web services and service level agreements. Message based communication is used in networking. Finally, coordination techniques are used in formal models of organizations to analyze or simulate them. In contrast, most coordination languages refer mostly to different kinds of metaphors, like blackboards, shared data-spaces, component composition and channels.

Multiagent systems consist of a set of agents, which can be designed and implemented in a variety of ways. In particular, a system designer would like to control the emergent behavior of the system. In human societies, groups of humans are controlled by organizational structures, for example in business. However, it is more difficult to define the relations between agents, or properties of the whole system. Therefore multiagent organizations are defined, which describe the relations between agents.

References

C.E. Alchourròn and E. Bulygin. *Normative Systems*. Springer, 1971.

G. Boella and L. van der Torre. A game theoretic approach to contracts in multiagent systems. *IEEE Trans. SMC, Part C*, to appear.

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.

J. Broersen, M. Dastani, and L. van der Torre. Beliefs, obligations, intentison and desires as components in agent architectures. *International Journal of Intelligent Systems*, to appear.

C. Castelfranchi. Modelling Social Action for AI Agents. *Artificial Intelligence*, 103:157 – 182, 1998.

C. Castelfranchi. Founding an agent's autonomy on dependence theory. In *Proceedings of ECAI'00*, pages 353 – 357. Berlin, 2000.

R. Conte and C. Castelfranchi. *Cognitive and social action*. UCL Press London, 1995.

E. Goffman. *The Presentation of Self in Everyday Life*. Doubleday, 1959.

J. Habermas. *The Theory of Communicative Action, Volume One, Reason and the Rationalization of Society*. Beacon Press, Boston, 1984. transl McCarthy, orig publ as Theorie des Kommunikativen Handels, 1981.

A. Jones and J. Carmo. Deontic logic and contrary-to-duties. In D. Gabbay, editor, *Handbook of Philosophical Logic*, page 203279. Kluwer, 2001.

J-J. Meyer and R. Wieringa, editors. *Deontic logic in computer science: normative system specification*. Wiley, 1993.

J.-J.Ch. Meyer. A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29:109–136, 1988.

S. Ossowski. *Co-ordination in Artificial Agent Societies*. Springer-Verlag, Berlin, 1999.

Y. Shoham and M. Tennenholtz. On the Synthesis of Useful Social Laws for Artificial Agent Societies (Preliminary Report). In *Proceedings of the National Conference on Artificial Intelligence*, pages 276–281, San Jose, CA, July 1992.

Göran Therborn. Back to Norms! On the Scope and Dynamics of Norms and Normative Action. *Current Sociology*, 50(6):863 – 880, 2002.

L. van der Torre. Contextual deontic logic: Normative agents, violations and independence. *Annals of Mathematics and Artificial Intelligence*, 37 (1-2):33–63, 2003.

H. Verhagen. *Norm Autonomous Agents*. PhD thesis, Department of System and Computer Sciences, The Royal Institute of Technology and Stockholm University, Sweden, 2000.

A Framework for the Design of Self-Regulation of Open Agent-based Electronic Marketplaces

Christian S. Hahn[⋆]
[⋆]German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, 66123 Saarbrücken
Germany
Christian.Hahn@dfki.de

Bettina Fley[*] and Michael Florian[*]
[*]Department of Technology Assessment
Technical University of Hamburg-Harburg
Schwarzenbergstr. 95, 21071 Hamburg
Germany
{bettina.fley,florian}@tu-harburg.de

Abstract

In this paper, we argue that allowing self-interested agents to activate social institutions during run-time can improve the robustness (i.e., stability, cooperation or fairness) of open Multiagent Systems (MAS). Referring to sociological theory, we consider institutions to be rules which have to be activated and adopted by the agent population. Informed by sociology, we propose a framework for self-regulation of MAS for the domain of electronic marketplaces. The framework consists of three different institutional forms that are defined by the mechanisms and instances that generate, change or safeguard them. We suggest that allowing autonomous agents both the reasoning about their compliance with a rule and the selection of the form of an institution helps to balance the trade-off between the autonomy of self-interested agents and the maintenance of ”social order” (cf. Castelfranchi (2000)) in MAS and to ensure almost the same qualities as in closed environments.[*]

1 Introduction

The design and development of open multiagent systems (MAS) where a vast amount of heterogeneous agents with different goals, different rationales and varying perceptions of appropriate behavior can interact, is an area of increasing importance in MAS-research, especially in the context of Internet applications (e.g., electronic marketplaces). Agent-based marketplaces consist of software agents who represent customers and providers, interacting with each other in order to trade goods and services (tasks), presumably acting on behalf of human users. In cases where providers cannot complete a task on their own, they also may be allowed to cooperate, form partnerships and organizations (cf. (Schillo et al., 2004) and (Schillo et al., 2002)). Therefore, designing *robust* and efficient open electronic marketplaces is a difficult challenge.

In accordance with Wooldridge et al. (1999), *robustness* is the ability of a system to maintain ”safety-responsibilities” even in case of disturbances. Relating to Schillo et al. (2001), robustness criteria regarding open agent-based electronic marketplaces are attributes like scalability, flexibility, resistance and agent drop-out safety that can be measured by the *relationship* between certain ”safety-responsibilities” (i.e., domain oriented performance criteria) and domain specific perturbation scenarios. Factors affecting those qualities can be divided into two groups:

Firstly, the model of the electronic marketplace itself can cause technical problems reducing the system's performance, since the technical feasibility of open MAS depends on agent interoperability, communication protocols, reliable infrastructures etc.

Secondly, environmental influences (e.g., demand for new products, newly participating provider agents) and interaction dynamics in the market can cause interaction outcomes and system states that are

[*]This work was funded by the Deutsche Forschungsgemeinschaft (DFG) in the priority program Socionics under contracts FI 420/4-3 and FL 336/1-3.

both undesirable from the perspective of user purposes (reliable and efficient electronic trade) and diminishing the trustworthiness and acceptance of such MAS applications. However, depending on the perturbation scenario, it may be not possible to precisely define ex ante those actions that lead to undesirable outcomes either on interaction or system level. The reason for this lack of an unambiguous definition of unwanted actions is that in economic contexts agents are usually not benevolent, but self-interested. Though the notion of markets implies that the pursuit of self-interest leads to efficient coordination, certain behaviors guided by self-interest may also cause market failure for several reasons. In this context, only deception and fraud (in contrast to the honest pursuit of self-interest) as a reason for inefficient market coordination can easily be defined as unwanted behavior. Other forms of self-interested behavior, which are actually essential in markets (e.g., adaptation of prices), can also cause externalities and forms of market failure, but only under certain conditions. Therefore, those behaviors cannot be declared as generally deviant. Finally, local knowledge and information asymmetries may lead to diverging goals and varying perceptions of options(plans) to fulfill the self-imposed goal, impairing the coherence of market interaction as well.

With our framework, we do not aim at providing solutions to all these factors affecting robustness, but at improving the robustness of open electronic marketplaces with respect to the following perturbation scenarios:

Failure of market interaction. Firstly, market exchange can fail due to unreliable infrastructures. In a large distributed environment like the Internet, unforeseeable states might cause an agent to drop out or a message to be delayed, reducing the quality of goods and services. Secondly, market exchange may fail due to the absence of a secure payment system. The payment between trading partners is not guaranteed, since no empowered instance monitors the payment system. Deceitful behavior is a third reason: the compliance with contracts of sales is often not guaranteed in electronic marketplaces.

Unfeasible Market Strategies. In human societies, agents try to improve their competitive positions depending on the structure of competition, their own positions and the demand situation. In our model, agents are able to pursue certain strategies, e.g., by forming organizations. Due to a lack of knowledge about the state of the market, they may choose unfeasible or inadequate strategies with respect to the state of the market and other agents' choices of strategies. These strategies may not only lead to a poor performance of the agents themselves, but to unacceptable system states.

Ruinous Competition. While the adaptation of prices is necessary to achieve market coordination, agents may choose to offer their services for dumping prices in order to receive orders in periods of low demand or to gain market shares of competitors. However, price dumping may lead to insolvency and market break down.

Monopolization and Cartelization. In our scenario, providers are able to form organizations (cf. (Schillo et al., 2004) and (Schillo et al., 2002)). Hence, powerful entities may try to build cartels or a monopoly on certain products in order to gain power, to set prices and to attain higher producer rents. While the formation of organizational forms improves robustness (cf. (Schillo et al., 2004)), a certain degree of market concentration reduces the efficiency and flexibility of the whole marketplace.

In Distributed Artificial Intelligence (DAI), a common approach to handle these kinds of problems is to resort to a trusted third party (cf. (Froomkin, 1997) or (Boella and Damiano, 2002)) that establish *conventions and norms* that standardize interactions, establish safeguards and guarantee that certain intended actions actually take place and unwanted situations are prevented. In the social sciences, these phenomena that structure and regulate the actions of human agents like customs, norms and laws are summed up by the term *"institution"*. As a common denominator, differing sociological theories define institutions as rules that are binding, because on the one hand they create meaning and guide the expectations of agents and on the other hand they possess a certain obligatory nature, claiming validity and prevalence (cf. (Esser, 2000)). The cited work in DAI mainly focuses on the third parties, i.e., *"instances"* (individual agents or corporate agents like organizations) that safeguard rules rather than on the rules themselves. However, this does not mean that the idea of institutions is extraneous in DAI. In contrast, the term institution has been used repeatedly in DAI, but it has been primarily employed as a sociological metaphor in the sense of social agreements upon interaction styles and shared meanings. State-of-the art research on electronic institutions largely deals with the institutionalization of transaction protocols (cf. (Colombetti et al., 2002), (Dignum, 2001), (Dignum, 2002) and (Esteva et al., 2001)). Some work (cf. (Axtell, 2001)) also uses the term institution synonymously with organizations, while ignoring that organizations are defined as social entities that consist of members, (material) resources and *rules* and that are able to act as corporative actor, whereas institutions are considered to be rules that are not capable of acting (cf. (Scott, 2001) and (Esser, 2000)).

Therefore, this paper aims at exploiting this metaphor from sociology more comprehensively in order to provide a theory of flexible institutions that allow agents in an electronic marketplace to activate *rules* (institutions) and to activate or form *instances* (third parties that generate and safeguard) during run-time in order to dynamically self-regulate non-desirable interactions and system states, using different mechanisms to gradually restrict the agent's autonomy.

2 A Sociological Perspective on Institutions

In sociology, institutions are defined as rules. However, in contrast to the notion of the use of the term ”rule” in computer science, this does not mean that institutions are clear definitions about what actions are allowed or forbidden and what are the consequences in case of rule violation. Explicit instructions may be an important attribute of certain institutions, but those are not only considered to provide stability of social life by constraining and guiding actions. The term rule and institution respectively also refers to social agreements about meanings, taken-for-granted assumptions and appropriate frameworks of action. Moreover, with respect to human societies, rules in the sense of instructions as well as in the sense of shared meanings and assumptions may be ambiguous, what allows agents a certain autonomy.

With respect to the purpose of this paper to improve the robustness of MAS, this firstly raises the question whereon the potential of institutions rests to generate shared assumptions, structure expectations and regulate the actions of agents, particularly because rules have no capacity to act like individuals (agents), collective and corporative agents (organizations, instances). And secondly, this involves the question how this potential can be used for the design of MAS-based open electronic marketplaces.

Giving an answer to the first question from a sociological perspective requires a more detailed definition of the term ”rule” and a more detailed explanation what causes the *binding nature* and *a certain claim of validity and prevalence* of these rules called institutions.

2.1 Definition of Institutions as Rules

In sociology, a variety of meanings of the term rule exists that can be summarized by two different conceptions of the word. On the one hand, a rule can be an underlying principle of action, defining which interactions are desired, which are unwanted or even forbidden in certain contexts and under certain conditions. Underlying principle means that a rule is available to the awareness of agents and can be more or less consciously mastered by them in the sense that agents may reflect about rule violation or rule-conform behavior of themselves or others. On the other hand, specific collective behaviors can show a rule-like character, i.e., certain regularity. The appearance of this observable regular pattern is not necessarily due to the existence of a rule as a guiding principle of action. It is simply produced by the aggregate of individual actions guided by the same structural constraints or environmental states (cf. (Bourdieu, 1990, 60f.)). Reasons for the resemblance of behaviors of certain agents are similar orientations towards a certain objective of action (goals or ”desires”), as well as shared meanings and assumptions (”plans”) about how certain things can be done under specific circumstances (cf. (Berger and Luckmann, 1966)).

2.2 Attributes of Institutions

(1) Although institutions are defined as rules, not any rule, i.e., not any observable or cognitive available pattern of behavior, is an institution. In sociology, institutions are considered to be social macro-phenomena. That means that institutions in contrast to sociological micro-level phenomena (i.e., face-to-face interactions between persons present), are *durable*. Moreover, their *scope in the social space* reaches beyond informal relationships and temporarily limited encounters among dyads or triads of individuals.

(2) This spatial and temporal scope is one reason for the *”transintentional” character of institutions*. Institutions as macro-social phenomena are considered to be existing largely independent of the will of agents. They are not being at the disposal of single agents, so that they appear as an external constraint. However, the externality of institutions is to a large extent caused by other factors: some authors (cf. (Berger and Luckmann, 1966)) argue that this partial independence of institutions from agents' intentions

is caused by mere routine and customization. Institutions provide agents with solutions to specific problems of action, i.e., shared meanings, knowledge, and patterns of how things can be done. The more those meanings and patterns become taken-for-granted certainties and are disseminated within a population, the more alternatives will be ignored or considered to be not feasible. However, other important factors for the *externality of institutions* are possibilities to impose sanctions on rule violating agents, either by collective moral disrespect or physical force. Especially those institutions that are available to the consciousness of agents facilitate discourses about which behaviors conform to and which violate certain rules.

(3) Despite the externality of institutions, agents' actions do not necessarily need to correspond to collective behavioral patterns. Agents still may violate rules. To explain why agents commit themselves to act according to rules, it is not sufficient to refer to external factors. In order to oblige agents, rules necessitate to be accepted as legitimate by the agents themselves. However, *commitment* can have varying origins. Firstly, agents may adopt a rule as a *taken for granted certainty* (cf. (Berger and Luckmann, 1966)). Secondly, agents may attribute a rule a certain *value of its own* (e.g., because a rule is meant to safe-guard collective goods or public welfare). Finally, the commitment may be due to *agents' own interests*, depending whether rule conformity or violating behavior is useful to reach one's goals or to avoid disadvantages (e.g., bad reputation, legal sanctions).

2.3 Three Forms of Institutions

This definition already suggests that diverse types of institutions can be distinguished, depending on the degree to which they are available to the awareness of agents, their socio-spatial scope, their degree of durability, externality and capability to commit agents, and in general the strength with which they claim prevalence and validity. Moreover, the previous section already anticipated that institutions are varying with respect to their capability to further, prevent or stop certain actions and hence, differ to the degree to which they restrict the autonomy of agents. In order to develop a framework for MAS that leaves as much autonomy as possible to agents, it is necessary to provide a typology of different institutions that identifies which type of institution is required for the regulation of the different, in the introduction specified non-desirable system states. Following Scott (2001), who discriminates three elements of institutions, we distinguish three types of institutions with regard to the degree to which they restrict the agents' autonomy and to which they provide solutions for the regulation of the perturbation scenarios mentioned. However, in the remainder of this paper, we mainly refer to the work of sociologist Pierre Bourdieu on rules, regular and regulated behavior (cf. (Bourdieu, 1990)) for two reasons: firstly, Bourdieu's habitus concept provides insights (cf. (Schillo et al., 2000)) that help to develop an agent model that enables self-interested agents to reason about their obedience to rules, so that they do not need to give up autonomy completely (cf. Section 3). Secondly, the field concept (cf. (Schillo et al., 2004), (Schillo et al., 2002)) provides a concept that allows analyzing driving forces of institutionalization processes, i.e. sources of institutional practice beyond individual actors (cf. Section 2.3). However, space restrictions do not allow a more detailed summary of those two concepts in this paper.

2.3.1 Practical Institutions (PI)

With this type, we refer to observable patterns of collective behavior that are not produced by the consciously managed obedience to a consciously available rule (cf. (Bourdieu, 1990, 60)). Instead, those patterns (or so-called strategies) result from the actions of agents (1) which try to accumulate different sorts of ”capital” that are accredited in a certain social context (field), e.g., reputation, economic profit, (2) which are confronted with the same structural constraints (similar competitive positions), and (3) which share similar dispositions of perception, reasoning and action (a similar habitus). Both, the generation and the durability of those institutions result to a large extent from the similarity of the agents' dispositions (habitus), which in turn have been acquired by agents through their long-term experiences in a certain social context (social field) and are conditioned by the similarity of competitive positions in those fields. Hence, the socio-spatial scope of those regularities is mainly restricted to a certain ”agent class”, i.e., to those agents sharing similar competitive position over a longer period and hence a similar habitus. As a consequence, the collective behavior of a specific class manifests itself in a certain collective style of action that is recognizable by other agents. This stylization of action confers externality and transintentionality to this rule-like behavior, effecting a certain claim of prevalence of this collective style, since it allows the classification of an agent, confronting it with class specific expectations of other competitors. Although stylization allows the recognition of a certain behavioral pattern as class specific, practical institutions are no formulated rules. Therefore, the agents' commitment is not influenced by possible sanctions. Rather, commitment towards a style is due to its feasibility with respect to the agents' goals and to the identification of an agent with a certain style, leading to conclusive actions regarding that style. With regard to open MAS, the advantage of this type of institution consists in providing different classes of agents with behavioral patterns of feasible actions ("plans") and hence, in contributing to the coherence of interaction between agents by learning practical strategies in the electronic marketplace.

2.3.2 Normative Institutions (NI)

With this type, we refer to rules that Bourdieu calls quasi-juridical principles (cf. (Bourdieu, 1990, 60)) and which also can be defined as norms that indicate behaviors that are acknowledged as honorable and morally approved. Norms are more or less formulated and consciously manageable . The socio-spatial scope of those norms is not restricted to specific agent classes, but to communities of agents who share certain values. Norms allow shared judgments of certain actions either as honorable or dishonorable/immoral. In contrast to classes, communities are not defined by the similarity of the agents' social positions, but by network-like relationships between agents that are characterized by trust and commitment towards each other. Therefore, the durability and validity of those institutions are caused by the stability of the relationships, which in turn are safe-guarded by sanctions imposed by members of the community on norm violators (collective disrespect and/or exclusion of norm violating agents). The commitment of agents to specific norms of a group can have several reasons: the interest of an agent to be member of a trustworthy network (i.e., social capital), to be acknowledged for honourable behavior (i.e., symbolic capital, reputation), while norm violation would lead to a loss of those kinds of capital. Moreover, commitments also may be due to the adoption of a norm as a certain value of its own or as an unreflected disposition of action (e.g., routine, habit). With regard to open MAS, the advantage of this type of institution is that undesirable actions and strategies of agents can be sanctioned. This refers to both: (1) clearly undesirable actions like fraud and (2) actions which are undesirable, but only problematic, if they occur on a large scale (bad quality, dumping prices). However, sanctions do not enforce norm conform behavior completely, but effect a loss of reputation and lead to exclusion of agents from the interaction in the community. Hence, the autonomy of agents is only affected partially, since they still may act norm violating.

2.3.3 Regulative Institutions (RI)

With this type, we refer to codified, formal law that has been brought up intentionally by the legislative in order to regulate certain social facts. Those law-like rules clearly indicate which actions are allowed or forbidden, and what are the possible consequences in the case of violation. The socio-spatial scope spans the whole legal room, i.e. the entire system. The transintentionality and externality of those rules are caused by procedures of legislation, jurisdiction and execution, while their validity can be enforced by sanctions. If the commitment of agents to law is not caused by the adoption (incorporation) of legal prohibitions and commandments as values of their own, sanctions in form of penalties create incentives to act rule conform. With regard to open MAS, the advantage of this type of institution is that those undesirable actions and strategies of agents that (1) are very harmful (deception), (2) exceed a certain level of occurrence, so that they can not longer be prevented or stopped by normative sanctions (e.g., price dumping), or (3) cannot be resolved by reputation or moral disrespect at all (e.g., monopolies) are regulated by law. However, regulative institutions signify a direct intervention in the agent's autonomy, especially in case that a "prison sentence" is imposed.

2.4 Mechanisms of Generating, Adapting and Reproducing Institutions

The description of the three types of institutions already provided some insights how institutions emerge and become prevalent, valid and binding. Rules can be generated in different ways. They either can be laid down intentionally by some social entity or they can emerge bottom-up through the repeated interactions between agents. Rules do not necessarily need to be formulated, established intentionally or to be codified to be valid. But in contrast, even if they are established by a single act (like laws), they need to be adopted and accepted by the agents and reproduced through their actions in order to be valid, because a certain degree of rule violation simply overextends the capability of the law monitoring instances to prosecute violations and enforce legal behavior by means of physical force. However, in order to develop a framework that allows agents to activate certain rules (i.e., certain contents that define appropriate behavior) and to adjust the type of the institutional form, it is necessary to explain the process how those different forms are generated, adapted and reproduced. According to Bourdieu (1990, 76 pp.), five general mechanisms involved in the process of institutionalization of any type of rule can be distinguished, while peculiar modes can be specified for each type (cf. Table 1).

2.4.1 Reflection

The generation and reproduction of rules as well as the change of an institutional type depends to a large extent on the mode of reflection about both the rule itself and the obedience to it. While the generation of practical institutions (PIs) does not mean that agents either need to intend to generate a collective style consciously or to act accordingly, the generation of normative institutions (NIs) necessitates discourses and reflections about which behaviors should be valuated in which way. Such discourses may happen in case that agents become aware that a former practical institution has become problematic, e.g., because new agents entered a field, so that they translate a pattern of action into an imperative for action, i.e., an assignable rule. If legislative actors or instances discover that some behavior that formerly has been ensured by the ”means” of practical or normative institutions becomes problematic, they may intentionally formulate a regulative institution (RI). Moreover, the reproduction of NIs and RIs by rule-conform behavior depends to some extent on the anticipation of the consequences of rules conform or violating behavior.

2.4.2 Formalization

The type of an institution changes depending on the degree to which a specific pattern of behavior is formulated as a rule. While PIs are hardly communicable, those patterns may exhibit inconsistencies and irregularities. NIs are more formalized and communicable. However, in contrast to law, which is intentionally established, formulated and codified and which is meant to be logically consistent, defining a corpus delicti precisely, NIs are more ambiguous and fuzzy.

2.4.3 Officialization

Institutions as sociological macro-phenomena have a certain socio-spacial scope, even though this may differ depending on the type of institution. In order to generate an institution or to change its type, it is necessary that either the assignable rules or the class-appropriate behaviors and strategies are diffused within a particular social space. While laws are made available to the entire population by publication, NIs are often spread and generated by gossip and denunciation of dishonorable behavior within a network. PIs itself are not communicable, however, agents of a certain class can comment the behavior of others regarding its feasibility and they may observe and imitate the behavior of others agents of their own class.

2.4.4 Objectivation

In order to be durable, any institution needs to be objectified, i.e., to become an objective fact, external to the will of agents. In the case of PIs, this is mainly achieved by the incorporation of the collective style into the dispositions of agents (their habitus or ”beliefs”), and by the transformation of a collective behavior into a taken-for-granted certainty. RIs are additionally objectified by forms of materialization, i.e. they are backed by material resources like courts, police etc. Also NIs can be objectified by material resources, e.g., certain monitoring associations.

2.4.5 Legitimation

The reproduction of a rule by conforming actions of agents, as well as the change of an institution into a type that restricts the autonomy of agents to a larger extent depends on the acceptance of that rule as legitimate. While PIs are perceived as legitimate by the agents of a class as long as they provide feasible strategies of action and serve the purposes of those agents, this legitimating reason is problematic with respect to NIs and RIs, since those institutions often try to prevent behavior that is rational from the perspective of single agents, but not from the perspective of the entire agent population. Therefore, those institutions are often legitimated by their contribution to common goods, welfare and public interests.

2.5 Sources of Institutional Dynamics

The question what are the driving forces that generate, reproduce and adapt institutions still remains, since institutions themselves are not capable of acting. While regulative institutions are established intentionally by agents, the other types of institutions are somehow induced less intendedly by agents who pursue their interests. Moreover, any type of institution needs to be reproduced by agents. However, this does not mean that only individuals are the driving forces of institutionalization processes. Also collective and corporative agents (i.e. groups or organizations that appear as single agents through their representation towards their social environment) can start institutionalization processes. This also does not mean that all agents pursue the same interests, e.g. legislative agents may be interested in the regulation

Mechanisms & Instances	Practical Institution	Normative Institution	Regulative Institution
Reflections	Generation: implicit Obedience: prereflexive	Generation: discoursive Obedience: $\pm reflexive$	Generation: intentional Obedience: reflexive
Formalization	-	Formulation	Codification
Officialization	Annotation of actions	Communication of rule Valuation of action	Publication Claim
Objectivation	Incorporation Self-evidence	Materialization Naturalization	Materialization
Legitimation	Feasibility Conclusivenss	Common Good Morality	Public Interest Public Welfare
Generating Instances	Instances	Reputation Networks Instances of Diffusion Associations	Legislative
Safe-Guarding Instances	-	Associations	Judiciary Executive authority

Table 1: Three forms of institutions: primary mechanisms of institutionalization and sources of institutional practice (instances)

of a market, because they need to legitimate their own existence by their contribution to further the public welfare. We call those driving forces of institutionalization processes that are not individual agents and that pursue other interest than economically thinking provider agents *instances*. With respect to our application scenario, we distinguish the following instances that generate and safe-guard the different institutional types:

Instances of Diffusion. In the economic field, consultants as well as specialist journals and newspapers contribute to the diffusion of feasible strategies (behavioral patterns or plans), valuations of honorable, morally approved behavior, and information about the current state of the market.

Reputation Networks. Another instance to diffuse valuations of honorable behavior are reputation networks, in which gossip is spread. These networks also improve the process of building models about competing agents in the market.

Associations. In markets, provider associations often play an important role in establishing normative institutions with respect to dishonorable providers and unfair competition (e.g., price dumping). Since those collective actors are often accredited, they are powerful regarding sanctions by communicating dishonorable behavior of certain agents. Moreover, they can sanction member agents by excluding them from the association in case of rule violation. An association provides secure trading conditions because of trustworthy trading partners. However, since associations take membership fees, they often provide additional incentives in order to acquire members.

State. With respect to RIs, the different corporate actors of the state, i.e. the relevant instances of legislation, jurisdiction and the executive, generate and safe-guard laws.

3 An extended BDI Architecture

As we have argued before, social phenomena like rules and obligations help to coordinate agents' interactions and thus to improve the agents' performance (cf. (Tennenholtz, 1998)), the regulation of e-commerce (cf. (Dignum, 1996)) and open electronic marketplaces (cf. (Dellarocas, 2001)). Although the improvement of agent performance sounds very promising, complying with rules and obligations is not always the most rational way to fulfill an agent's goal. In some cases, rules directly conflict with the agent's desire. Or the compliance of two or more rules is impossible, since the action needed to fulfill a single norm violates other rules. Therefore, the agent must be able to rationally decide which actions or plan to choose in order to comply with institutionalized rules and to reach the self-imposed goal. So, when an agent considers which course of action to follow, before it takes a decision, it depicts in its understanding the consequences of its action in terms of its own welfare and the expectations and preferences of the remaining agents in the society expressed by an obligatory rule.

Various competing architectures for MAS have been proposed, and it is still unclear which type of architecture should be used for which type of application. However, with respect to economic application domains, agents should be able to interact with other agents by autonomously selecting an action on the basis of their own interests and goals (profitable trade, accumulation of social or symbolic capital). In this context, the most common agent architecture is the BDI-architecture (cf. (Bratman, 1987)) that bases on beliefs, desires and intentions. However, in this architecture social concepts like institutions, i.e., obligations and rules are not considered.

In order to leave as much autonomy to the agents as possible, the integration of autonomous agents into societies regulated by institutions demands an architecture that includes some facility to reason about complying with a rule and about the consequences of the subsequent actions with respect to the goals of an agent. Therefore, the agent must know both: which actions or plans to choose in order to comply with an institution and which to choose in order to reach its self-imposed goal. So, when an agent reasons about the selection of a course of the developed plans before making a decision, it evaluates the consequences of the available options of action, based on the utility of certain actions with regard to the accumulation of different sorts of capital like economic capital or symbolic capital (e.g., reputation) and the effects of rule violation or compliance on that utility. Nevertheless, as we have argued in Section 2.2, rule complying actions can firstly be carried out implicitly and prereflexive, based on the agents' dispositions (habitus). Secondly, they can be a consequence of reflections or reasoning about negative incentives (sanctions). But note that in case of sanction-based institutions, the compliance with them can also provide additional ”utility”, i.e., incentives like reputation for honorable behavior (symbolic capital) or advantages due to the affiliation in a network (social capital). Moreover, agents may attribute a value to some rules for its own sake and incorporate this rule into their own dispositions (beliefs and plans). This means that the architecture must provide two modes to select actions: (1) institution-reflecting pursue of interests (desires) and (2) disposition-based rule compliance in the sense that those rules are taken for granted and no alternative (rule violation) is taken into consideration. Here, the behavior, which is executed as a consequence of the agents' dispositions, is internally represented by the constellation of the agent's beliefs, desires and intentions, but not recognized as a sort of institution. In contrast, institution reflecting actions require the explicit representation of a rule in the internal structure of the agent. Hence, in order to allow different attitudes towards institutions, we have to provide some basic functionalities. Firstly, the agents should be able to *recognize* the existing rules (NIs, RIs) on the one hand and to *adopt* certain behavioral patterns (PIs) or rules (NIs, RIs) into their beliefs, desires and intentions on the other hand. Secondly, the agents need to be provided with a kind of reasoning mechanism allowing them to decide if certain dispositions or rules should be adopted, or if rules should be deliberatively followed or violated (in case of NIs and RIs). Thirdly, if the agent adopted a rule, it should be able to react on deviant behavior by using the sanctions corresponding to a rule. In the following, we present a BDI approach that allows to meet these requirements.

3.1 The $R \subset BDI$ Architecture in Theory

Disposition-based agents do not regard their behavior as rule complying, because the dispositions towards rule conform actions are already internally represented. The BDI-architecture, a simple yet powerful and elaborated agent model (cf. (Bratman, 1987)), provides in this context an option to realize this representation by translating agent dispositions into beliefs, desires and intentions, the three components of the general BDI concept:

- Beliefs (B) represent the agent's mental model of the world (the market) including the consequences of its action. Regarding our application scenario (electronic marketplace), these beliefs include the agent's current knowledge about the structure of the economic field or market (competitive position of other providers), the demand situation and its own position as well as knowledge about achievable profits (economic, cultural, social, and symbolic capital).
- Desires (D) reflect a projection, a goal the agent wants to pursue. Regarding our application scenario, these goals can vary with respect to the sort of capital an agent wants to accumulate.
- Intentions (I) translate a goal into actions. They are future directed and often lead to the execution of a plan. In our scenario, these intentions are feasible competitive strategies, consisting of a combination of possible courses of a number of actions (plans).

From out point of view, there are three advantages of the BDI model: (1) it allows to build agents that shows different behaviors within a single architecture. (2) The basic BDI deliberation cycle provides for a spectrum of behaviors ranging from purely deliberative to highly reactive, depending on the structure of the generated plans and on the decision, how often to take changes in the environment into account while executing plans. (3) BDI-agents are able to operate in dynamic environments and to flexibly and appropriately adapt to changing circumstances despite incomplete information about the state of the world and other agents in it. However, this model allows only disposition-based rule compliance, even if the agent is provided with rule-conform beliefs, i.e., obligations to execute some actions within a given set of agents. Though, agents may adopt a rule, adoption means that a rule is implicitly positioned in the beliefs of the agent and that corresponding desires are formulated so that a rule-driven action can be executed. This has the effect that agents always will comply with a rule once they adopted it, even in absence of sanctions and under ignorance of their original interests. With respect to our application scenario, this can be quite problematic, because users would not accept applications where autonomous agents acting on behalf of them ”forget” their own purpose (i.e. achieving economic profits by trade) too easily. Therefore and in order to integrate them in our framework, it is necessary that they do not automatically adopt a rule and comply with it. This means that they firstly need to be capable to reason, if they should comply with a rule with respect to their goals and the current market environment. And secondly, they need to be able to reason whether an adoption of the rule is useful, depending either if there are alternative intentions which are both rule-conform and compatible with the agent’s desires or if a certain rule helps to achieve a goal (e.g. prevent price-dumping) that otherwise is not realizable. So, the extension of the known BDI model integrates two aspects: *rules* and a *utility* function.

- Rules R reflect the current institutional rules and possible sanctions related to that institution. These rules are perceived in the environment, illustrating a subset of the current Bs.

- Plan Repository PR stores feasible plans that the agent could follow in order to achieve a self-imposed goal.

- Function f describes a utility function which evaluates the utility of feasible plans. This function bases on experiences with certain strategies and plans and related goals in the past.

In the next section, the extended BDI-architecture is presented. Aiming at balancing the agent’s intentions and the social rules, the architecture is confronted with the problem how an agent may decide which rules are compatible with its intentions (strategies) and desires. Especially NIs and RIs, which are often legitimated by common goods and public welfare, represent common interests (desirable system states) that easily conflict with the intentions or even with the desires of single agents. However, they also provide incentives (sanctions/ rewards) that enable to balance personal and public interests. According to (Conte and Castelfranchi, 2001), a social incentive (positive or negative) is an additional value or utility that modifies the agent’s decision and is provided by an external entity. In order to enable reasoning about those incentives, rules and incentives must be explicitly represented in the set of beliefs, since the reasons to comply with a rule may change (e.g. avoiding sanctions, achieving rewards or even a goal that presupposes a certain rule).

3.2 The $R \subset BDI$ Architecture in Practice

The $R \subset BDI$ architecture mainly differs regarding the explicit representation of rules and the associated sanctions, rewards and the safe-guarding instances of the rule. The reasoning process described in this section deals with the question why an agent should accept an institution as a rule. Once an institution has been recognized as a rule, an additional belief is formed, representing the rule itself and possible consequences (sanctions/rewards). But the recognition of these rules is not sufficient to initiate the formation of a new goal. A rule is conformed, if the agent sees it as a way to achieve the superior goal, i.e., if the rule is instrumental for achieving the main goal. In general, the goal of each agent is described by the improvement of the market position in the economic field. Two subgoals are instrumental for reaching this goal: firstly, the improvement of the own position and secondly, the degradation of rival agents’ market position. The improvement of the own position can be reached either by the reduction of uncertainties inside the market (e.g., the activation of institutions) or the selection of a practical market strategy (e.g., cost reduction, diversification or monopolization etc.), which in turn reduces the position of rival agents. The second possibility to influence the market of competitors is to practice deviant behavior, e.g., by the diffusion of wrong information. Figure 1 shows the $R \subset BDI$ approach and the internal processes to select a goal. The revision of goals is triggered by events like a change of market information (leading to an alteration of the agent's market strategies and, therefore, the agent's goals). In that case, new plans have to be created on the basis of the possible new goal. PR comprehends all possible plans the agent could develop so far in order to reach the set of feasible goals, where the selection of an adequate plan to meet the goal is left to the agent itself. If the set of new Bs does not drastically influence the agent's decision, the goal is kept at least until the experiences about the last strategy are updated in the reasoning process.

The core element of this architecture consists of the reasoning functionality of the $R \subset BDI$ agent. Usually the reasoning process is designed as follows: the set of B establishes the basis for the formulation of the desires, while a utility function is the decisive factor for the selection of the plan out of PR that has the ”maximum” potential to meet the defined main goal from the perspective of current beliefs. The utility function f maps a real number to each action or state expressing the estimated utility of that market strategy. In general, the plan $P_j(S)$ is the one which is considered to further the achievement of own goals to the largest extent.

$$f(P_i(S)) \leq f(P_j(S)), \;\; \forall P_j \in PR \wedge P_i \in PR$$

This equation is easy to evaluate for scenarios of total certainty where agents can reach their desires by executing a single action. However, the execution of several actions leads to the problem of an increasing probability of reaching uncertain system states because of each agent's socially bounded rationality (cf. (Schillo et al., 2000)) to estimate future events. For instance, the effects of sanctions and rewards on the basis of reputation and gossip can hardly be estimated. In principle, an agent must evaluate its own actions on the basis of the possible reactions of the other agents. However these are unknown in open electronic marketplaces that consist of heterogeneous agents. Thus, we have to redesign the function f, by introducing an expected utility (cf. (Parsons and Wooldridge, 2002)) and maximizing[1] the function f.

$$P = \max \sum f(P_i(s_n)) \; pr(s_m, a_i, s_n),$$
$$where \, s_m, s_n \in S \wedge P, P_i \in PR \wedge a_i \in A \wedge a_i \in P_i$$

In principle, this equation produces the utility value of the apparently most practical plan and expresses the uncertainty by a probability distribution so that plan P_i will transfer the agent from state s_n to s_m. Sanctions and rewards are complex factors for reasoning, not least since two sorts of sanctions can be distinguished: (1) sanctions of which the effect of the deviant behavior is known before an action is executed (in case of RIs), and (2) sanctions of which the kind of sanction is known, but not the exact consequences (e.g. effects of reputation). The inclusion of the first sort of sanctions into the evaluation of the expected utility is not difficult, whereas the utility of deviant behavior of the second type can hardly be estimated. In order to solve this problem of evaluating the expected utility, a reinforcement learning strategy is introduced. Q-learning (cf. (Watkins, 1989)) is a recent form of a reinforcement learning algorithm that does not need a model of its environment and can be used on-line. Nevertheless, the environment is taken into consideration, since Q-learning algorithms work by estimating the values of state-action pairs and states may represent environmental factors (e.g., competitors). Therefore, it is very suited for repeated interactions with unknown competitors. The value $Q(s, a)$ is defined to be the expected discounted sum of future pay-offs that is obtained by taking action a from state s and following an optimal policy thereafter. Once these values have been learned, the action that is considered to be the most practical from any state is the one with the highest Q-value. While using arbitrary numbers when initialized, Q-values are subsequently adapted on the basis of experiences. Space restrictions do not allow a detailed description of the evaluation of the best strategies (plans) by Q-learning. However, it should be mentioned that a new element has been added to the algorithm: an additional utility value for the known sanction and reward effects. If those values are explicitly known, they are not included in $f(s, a)$. In contrast, the unknown values of good or bad reputation are included in the function $f(s, a)$, since they can only be learned by monitoring the future course after executing action a.

So, a new goal in D is selected, if a market strategy shows higher utility regarding the improvement of the market position than the selected strategy in the past. Reasons why the new strategy is predicted to be more useful than a former strategy are sanctions in case of violating an institutionalized rule. We stated before that an institution is not sufficient for the formation of a new goal. So, another ingredient is needed, namely possible consequences caused by rule violation, reducing the strategy's utility. Obviously, if the agent is not able to perform the strategy that conforms to the institution (cf. Figure 1, in the component *intention*)

[1] Note that the term ”maximizing” here is not used in the sense of economic theory where utility maximization is restricted to perfect information and rationality. Maximization here means mapping plans and states on the basis of current beliefs about the social environment that are represented in the agent's mental model.

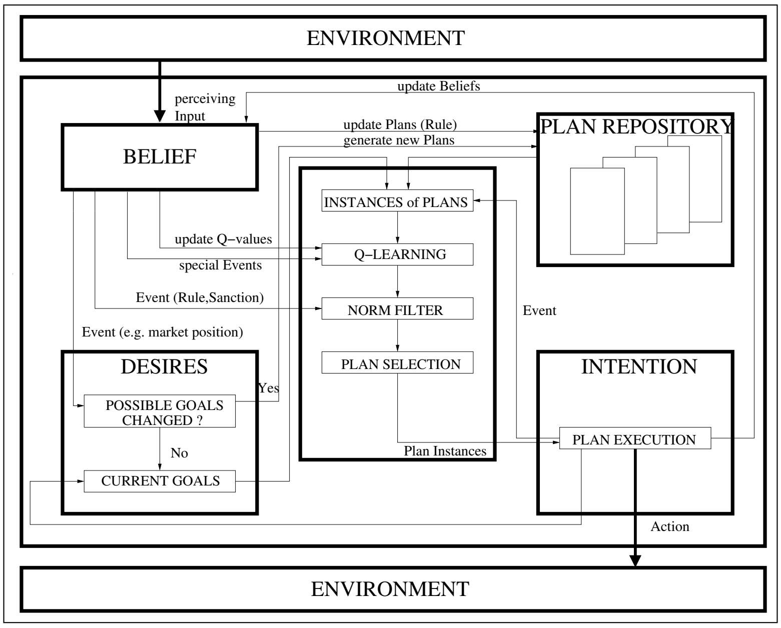


Figure 1: The $R \subset BDI$ architecture.

although the rule has been recognized and accepted, the agent has to find the next best strategy, hoping that it is able to execute the necessary actions.

4 Specifications of Institutional Forms

In the previous two sections, we defined social institutions as rules that vary with respect to the reasons to comply with them or to adopt them. Depending on the type of a rule (practical, normative or regulative), institutions are able to define meanings and to structure and regulate the interactions and the behavior between the agents to different degrees. However, since institutions are defined as social rules and not as social entities, they are not capable of action themselves so that they neither can safe-guard themselves nor ”change” their type. Institutions depend on individual, collective and corporative actors to be existent and valid in order to constrain and guide the agents’ actions. In this section, we discuss possibilities how agents can generate, adapt and change institutional forms in order to self-regulate their behavior and undesirable system states. In order to specify the self-regulation process, it is necessary to distinguish between the (1) content of a rule and the (2) form of a rule. The content specifies the behavior that is typical for a certain collective style (PIs), that is honorable (NIs) or that is forbidden (RIs). However, same contents can be reproduced and safe-guarded by different institutional forms which are defined by the mechanisms and instances that generate, adapt and change them. We assume that rules are represented as soft constraints that enable the detection and sanctioning of violations and not as hard constraints that are designed to make such violations impossible. Before explaining the process of self-regulation, the different collective and corporative actors (instances) that generate, safe-guard and reproduce rules are introduced.

4.1 Instances of Regulative Institutions

A juridical and an executive power guarantee that the regulative institutions (law) are met by the provider population. Deviant behavior can firstly be monitored by the executive power and, secondly, provider agents that recognize such behavior are able to bring that deviant provider to court. The court decides if the deviant behavior should be punished on the basis of evidence. Both, the juridical and the executive power are modeled as autonomous agents. The sanctions base on economic capital and are published when the rule (regulative institution) is passed. This accessible information is considered when a new future plan is evaluated in order to meet the agents’ desires. In general, the decision to violate an adopted right bases on the estimation whether deviant behavior is preferable with respect to the utility that is influenced by sanction dues and achievable rewards.

4.2 Instances of Normative Institutions

4.2.1 Journalists and Consultants

Instances of this institutional form can be journalists and consultants which spread information about the behavior and the reputation of provider agents.

4.2.2 Reputation Networks

In case that a normative institution could not be established, because no shared definitions of deviant behavior exist among provider agents yet, providers may spread gossip about agents to indicate if they judge certain actions and certain agents as honorable. This form of denunciation may also be spread by the journalists to support the formation of a public opinion. Negative reputation has drastical effects on the performance of provider agents, since the customer population decides to which agent a job is assigned on the basis of the agents' reputation.

4.2.3 Associations

Agents that are either affected by denunciation or by deviant behavior are interested in establishing explicit rules that define more precisely which kind of behavior is deemed as deviant. Therefore, associations can be formed in order to define explicit norms according to which each member is obliged to act and behave. Deviant provider agents which are members of an association are excluded. The scope of each association is restricted to its members. For this reason, the association has not the power to directly punish nonmembers, but indirectly by gossip as a kind of denunciation which is diffused by the journalists. However, an association is not only a simple control instance to ensure the standardization of rules. It also offers services to its members by representing them and by providing an exception handling mechanism in order to create incentives to acquire members.

Firstly, the association represents its members by publishing the creation purpose, so that customer agents could identify an association and its members. As we mentioned before, the probability of perturbations (e.g. agent drop-outs) increases in open MAS. The acceptance and legitimation of the institution that has been produced by an association on the customer side are therefore necessary for economic reasons, because negative reputation affects the performance of the members of the association in terms of task allocations. Therefore, the association members select the most trustworthy member as representative to express and further the trustworthiness of their association. Since the furtherance and maintenance of the association's good reputation are two of its main purposes, only reliable provider agents are allowed to join.

Secondly, an exception handling mechanism as a special sort of service offered by the association allows to increase the flexibility, resistance, reliability and drop-out safety. Common MAS research favors an individualistic approach to handle exceptions. This means that all failures that happen in the direct area of an agent are recovered by itself. Especially in open systems, this approach can not guarantee the expected quality and behavior for some reasons. The single agents do not have a global view of the current market state which is notoriously difficult to create without heavy bandwidth requirements. In contrast, we favor a semi-centralistic approach by authorizing the association to negotiate between members to prevent failures. In case of an exception like an agent drop-out, deadline passing or deceitful behavior, the institution intervenes and coordinates the ongoing agent interactions. For this purpose, member agents which were not able to get all their free resource allocated could register their free resources at the association. Note, the decision to enlist is unsolicited, so that the autonomy of members is not restricted. Moreover, if an agent has got a job assigned in the past, but is not able to fulfill the requested task in the present (e.g., due to an agent drop-out), the agent could request the association, if it is able to negotiate between itself and the registered agents. This exception handling mechanism protects the reliability of the association and improves the performance of the involved agents. From our point of view, this institutional approach is more powerful than any individualistic or organizational approach in providing a recovery system, since both the individualistic and organizational approaches do not possess enough authority and capacities to ensure a comparable performance.

4.3 Instances of Practical Institutions

The actions of individuals are not only guided by explicit rules and obligation, but also by conceptions. In the case of practical institutions, journalists and consultants diffuse information about profitable strategies (plans), while agents whose strategies could not convince may adopt them. In fact, the agents' commitment here is due to its will to ”survive” economically, since losses can only be compensated by profitable strategies before the agent gets insolvent and is excluded from the e-market. Surely, the total provider agent population will not adapt the same strategies, since the feasibility of strategies depends on competitive positions. The formation of a practical rule depends on the agent itself. Agents having adopted the suggested actions build an internal formalization and representation of that rule. The other agents, who have not adopted the institution, do not incorporate any representation, even if they share the same disposition of that action. In special cases, the institutionalized strategy could possibly be propagated by gossip. The legitimation of that practical institution and its motives bases on the acceptance of both the agents that activate and reproduce the institution and their counterparts.

5 A Self-regulation Process

As already mentioned, the hypothesis we propose in this paper is that the process of institutionalization can increase the robustness, efficiency and performance of open MAS. However, we also argue that maintaining the same form of an institution statically is not appropriate in order to handle dynamic disturbances and to close the gap between both providing each agent with sufficient autonomy and to ensure the system qualities. To adequately model self-regulation processes, we have to provide the system with two possibilities: the activation and the resolution of institutional forms.

5.1 The Activation of Institutional Forms

We argue that allowing self-interested agents to dynamically activate an adequate type of institution improves the system's performance in case of environmental changes. After careful considerations, the question how to ”change” an institutional form turns out to be an open research question. The process to find an appropriate order according to which the institutional forms should be changed is complicated, since the different institutional forms do not necessarily presuppose the existence of each other. From a sociological point of view, the arrangement of the institutions on a spectrum is inappropriate, since an institutional rule (a certain content) does not necessarily pass through the single institutional forms. The normative institution should not be seen as the consequence of a practical institution and a regulative not as the consequence of a normative one. Therefore, we decide to consider the process of activating a certain institutional form isolated from the two other forms and let the agents themselves decide which form they want to activate. For this purpose, we recapitulate the preconditions of the emergence of institutions.

5.1.1 The Activation of Practical Institutions

Certain types of actors develop habits or patterns through interaction. Institutionalization occurs as these interactions are reproduced and become taken for granted and internalized as goal so that other alternatives are not be recognized as potential actions. However, these institutionalized actions may not be the most profitable because of the restricted capabilites and missing experiences of agents about the behavior of other agents and the market. In general, three possibilities to activate practical institutions exist: (1) experience, and (2) gossip as well as (3) consultants and journalists. The process of adapting behavioral patterns is triggered by the publication of practical market strategies by journalist agents. These strategies have already proven their feasibility in previous market interactions. The agent recognizes this strategy and compares the utility of that strategy with the utility of its strategies used and developed in the past. If the published plan shows a higher utility, a new strategy is adapted, changing the current goal. Otherwise, the published strategy is ignored and the agent keeps on going to achieve its ”old” goal.

5.1.2 The Activation of Normative Institutions

As already mentioned, a normative institution is activated by agents that are interested in common and explicit norms in order to escape from illegitimate denunciation or deviant behavior of others. Denunciation (bad reputation) leads to a drop of orders. Note that the agent should assure, whether the order drop is a consequence of its own bad market strategies, of deviant behavior like a breach of contract or of mere denunciation. Depending on the causes of the order inclination, strategies like gift-exchange in order to strengthen the relationships to customer agents may be more appropriate than activating institutions. But if the conditions to activate an institution are fulfilled, the creation of an association that protects and ensures the compliance with predefined rules is stimulated. For the membership in such an association monthly dues have to be paid. So, agents that demand fixed rules should have good reasons, when they intend to establish predefined rules. An agent has to be massively aggrieved in order to accept additional costs. In most cases, possible rewards are not completely and precisely known so that they can not support the decision making. Therefore, the agent has to balance the negative effects of missing rules and the steady association costs in order to come to a decision in the reasoning process.

5.1.3 The Activation of Regulative Institutions

This form of institution is activated by an agent that feels aggrieved and, thus, enforces a claim by legal actions. Both types of agents (provider and customers) could ask for such a claim, since deviant behavior affects provider and customer agents to the same extent. The reasoning process is very similar to that of normative institutions. Additionally, the agent should have the belief that reputation as sanctioning mechanism (as in the case of normative institutions) does not have the potential to cope with this form of perturbation. However, an intuition which form of institutions might be appropriate could only be achieved by experiences with the different sorts of perturbation.

The claim that agents may enforce has to correspond with the system designer's normative concepts. Moreover, in order to avoid that the agents go to court to prevent undesirable system states whenever an undesirable action is carried out, each claim is assigned with costs that the agent has to pay in order to have the possibility to activate a regulative institution. So, the action to build a regulative institution is only selected as the best strategy by the Q-learning function, if it shows the highest utility of all feasible strategies. Consequently, this sort of action is selected as the last resort.

In principle, we take up the argument of (Conte and Castelfranchi, 2001) in our arrangement of the institutional forms. Castelfranchi defines social order as a pattern of interactions among interfering agents that allows the satisfaction of the interests of some agents. In this section, we have presented the activation process of institutional forms. We have argued that the institutional approaches can flexibly be arranged, i.e., the agents do not have to follow a predefined spectrum. Nevertheless, in order to model a self-regulation process, we should allow the agents also to ”resolve” an institutional form themselves, if certain perturbations do not occur anymore.

5.2 The Resolution of Institutional Forms

In sociological theory, an institution is durable meaning that it could not easily be resolved. With respect to MAS, the resolution of institutions makes sense in case that the disturbance that has been the source of the activation of an institution does not longer occur. Even though this appears reasonable, we suggest that the resolution of institutions is only useful in particular situations, depending how much they restrict the agent's autonomy. Practical institutions should not and cannot be resolved at all, since those behavioral patterns are constitutive for any interaction. Nevertheless, since practical institutions are changing continuously as the market situation changes, this is no problem. In addition, they assist the provider population to improve their performance. Normative institutions are activated in order to create a clear and definite norm, expressing a common standard. However, the purpose of an explicit norm is reached when the population meets this standard. In this case, the resolution of a normative institution is conceivable and the resolution of an association should be instituted by the representative of that association. In contrast to practical and normative institutions, the option to cancel a regulative institution affects the autonomy of the provider agents drastically. Therefore, in the case that each provider agent acts according to the institutionalized law, the reason to establish a regulative institution ceases to exist and thereby the reason why the agent's autonomy should be restricted further on. The resolution process is invoked by the legislative power. In the case that the same disturbance occurs after the resolution, the institutional form is recalled to life by the legislative power and exists then for a predefined duration. This means that although the regulative institutional is canceled and the instance has no right to sanction, it can be activated again in the future.

6 Conclusions and Future Work

In this paper, we argue that allowing agents autonomously to activate rules (institutions) and to self-regulate the form of institution may have interesting and useful effects on the performance of open electronic marketplaces. We have seen that social institutions are more than a pure rule system. Associations, journalists and an executive power and other instances as well as single agents ensure the activation and compliance. Especially these instances influence the behavior and performance of the institutional members in different ways. Associations, for instance, provide a recovery system allowing to increase the fault tolerance. Each of the three institutional forms is only to a certain degree appropriate for the different perturbations scenarios sketched in the introduction, since they affect the agent's autonomy differently. The self-regulation process allowing the agents to activate institutional forms during runtime presents therefore an appropriate mechanism to increase the system performance and the robustness of the open MAS. The resolution of institutions permits to efficiently react on changing environmental states. However, these assumptions about the benefits of our framework have to be tested by experiments. Currently, a prototype of the system is implemented.

Acknowledgments

We are indebted to Dr. K. Fischer, Dr. F. Hillebrandt, D. Spresny and I. Zinnikus for their support and many constructive ideas and comments.

References

R. Axtell. Effects of interaction topology and activation regime in several multi-agent systems. In *Multi-Agent Based Simulation: Second International Workshop on Multi-Agent Based Simulations Boston, MA USA, July 2000*, volume 1979 of *Lecture Notes in Artificial Intelligence*. Springer, Berlin, Heidelberg and New York, 2001.

P. L. Berger and T. Luckmann. *The Social Construction of Reality*. Doubleday, New York, 1966.

G. Boella and R. Damiano. A game-theoretic model of third-party agents for enforcing obligations in transactions. In *Proceedings of LEA 2002 workshop*, 2002.

P. Bourdieu. *In Other Words. Essays Towards a Reflexive Sociology*. University Press, Polity Press, Stanford/ Cal, Cambridge/UK,, 1990.

M. Bratman. *Intentions, Plans and Practical Reason*. Harvard University Press, 1987.

C. Castelfranchi. Engineering social order. In *Proceedings of ESAW'00*, Berlin, 2000.

M. Colombetti, N. Fornara, and M. Verdicchio. The role of institutions in multiagent systems. In *Atti del VII convegno dell'Associazione italiana per l'intelligenza artificiale (AI*IA 02)*, Siena, 2002.

R. Conte and C. Castelfranchi. Are incentives good enough to achieve (info)social order. 2001.

C. Dellarocas. Negotiated shared context and social control in open multiagent systems. In *Social Order in Multiagent Systems*. Kluwer Academic Publishers, Boston, 2001.

F. Dignum. Autonomous agents and social norms. In *ICMAS'96 Workshop on Norms, Obligations and Conventions*, 1996.

F. Dignum. Agents, markets, institutions and protocols. In F. Dignum and C. Sierra, editors, *AgentLink*, volume 1991 of *Lecture Notes in Computer Science*. Springer, 2001. ISBN 3-540-41671-4.

F. Dignum. Abstract norms and electronic institutions. In G. Lindemann, D. Moldt, M. Paolucci, and B. Yu, editors, *Proceedings of Regulated Agent-Based Social Systems*, volume 2934 of *Lecture Notes in Artificial Intelligence*, pages 93–103, Berlin, Heidelberg and New York, 2002.

H. Esser. *Soziologie. Spezielle Grundlagen Bd. 5: Institutionen*. Campus, Frankfurt a.M., 2000.

M. Esteva, J. A. Rodriguez, C. Sierra, P. Garcia, and J. L. Arcos. Agent-mediated electronic commerce (The European AgentLink Perspective). volume 1991 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg and New York, 2001.

A. M. Froomkin. The essential role of trusted third parties in electronic commerce. pages 119–176, 1997.

S. Parsons and M. Wooldridge. An introduction to game theory and decision theory. In S. Parsons, P. Gmytrasiewicz, and M. Wooldridge, editors, *Game Theory and Decision Theory in agent-based Systems*. Kluwer Academic Publishers, Boston, 2002.

M. Schillo, H-J. Bürckert, K. Fischer, and M. Klusch. Towards a definition of robustness for market-style open multi-agent systems. In *Proceedings of the Fifth International Conference on Autonomous Agents*, pages 75–76, 2001.

M. Schillo, K. Fischer, B. Fley, M. Florian, F. Hillebrandt, and D. Spresny. FORM - A sociologically founded framework for designing self-organization of multiagent systems. In G. Lindemann, D. Moldt, M. Paolucci, and B. Yu, editors, *Regulated Agent-Based Social Systems*, volume 2934 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 156–175, Berlin, Heidelberg and New York, 2002. Springer.

M. Schillo, K. Fischer, F. Hillebrandt, M. Florian, and A. Dederichs. Bounded social rationality: Modelling self-organization and adaption using habitus-field theory. In *Proceedings of Modelling Artificial Societies and Hybrid Organisations*, pages 112–122, 2000.

M. Schillo, T. Knabe, and K. Fischer. Autonomy comes at a price: Performance and robustness of multiagent organizations. In F. Hillebrandt and M. Florian, editors, *Adaption und Lernen in und von Organisationen*. Westdeutscher Verlag, Wiesbaden, 2004.

R. W. Scott. *Institutions and Organizations*. Sage Publications, Thousand Oaks, 2nd edition, 2001.

M. Tennenholtz. On stable social laws and qualitative equilibria. *Artificial Intelligence*, 120(1):1–20, 1998.

C. Watkins. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, 1989.

M. Wooldridge, N. Jennings, and D. Kinny. A methodology for agent-oriented analysis and design. In *Proceedings of the 3rd International Conference on Autonomous Agents, AA99*, pages 69–76, 1999.

A Normative Framework for Agent-Based Systems

Fabiola López y López[⋆]
[⋆]University of Puebla
México
fabiola@cs.buap.mx

Michael Luck[†]
[†]University of Southampton
United Kingdom
mml@ecs.soton.ac.uk

Mark d'Inverno[‡]
[‡]University of Westminster
United Kingdom
dinverm@westminster.ac.uk

Abstract

One of the key issues in the computational representation of *open societies* relates to the introduction of *norms* that help to cope with the heterogeneity, the autonomy and the diversity of interests among their members. Research regarding this issue presents two omissions. One is the lack of a canonical model of norms that facilitates their implementation, and that allows us to describe the processes of reasoning about norms. The other refers to considering, in the model of normative multi-agent systems, the perspective of individual agents and what they might need to effectively reason about the society in which they participate. Both are the concerns of this paper, and the main objective is to present a formal normative framework for agent-based systems.

1 Introduction

Norms have long been used as mechanisms to limit human autonomy in such a way that coexistence between self-interested and untrusted people has been made possible. They are indispensable to overcome problems of coordination of large, complex and heterogeneous systems where total and direct social control cannot be exerted. From this experience, the introduction of *norms* that help to cope with the heterogeneity, the autonomy and the diversity of interests among agents has been considered as a key issue towards the computational representation of *open societies* of agents (Luck et al., 2003).

Although efforts have been made to describe and define the different types of norms that agents have to deal with (Dignum, 1999; Singh, 1999), work has not led into a model that facilitates the computational representation of any kind of norm. Each kind of norm appears to be different, which also suggests that different processes of reasoning should be proposed. There are some work that introduces norms in systems of agents to represent societies, institutions and organisations (Dellarocas and Klein, 2001; Dignum and Dignum, 2001; Esteva et al., 2001; Shoham and Tennenholtz, 1995). This research is primarily focused at the level of multi-agent systems, where norms represent the means to achieve coordination among their members. There, agents are assumed to be able to comply with norms, to adopt new norms, and to obey the authorities of the system but nothing is said about the reasons why agents might be willing to adopt and comply with norms, nor about how agents can identify situations in which an authority's orders are beyond its responsibilities. That is, although agents in such systems are said to be autonomous, their models of norms and systems regulated by norms do not offer the means to explain why *autonomous* agents that are working to satisfy their own goals, still comply with their social responsibilities. In addition, although the importance of modelling compliance with norms as an autonomous decision has been identified by several researchers (Castelfranchi et al., 2000; Conte et al., 1999a; Conte and Dellarocas, 2001; Conte et al., 1999), the issue is only partly addressed by others whose proposals for norm compliance generally rely on specific decision-making strategies based on how much an agent gains or loses by complying with (Barbuceanu et al., 1999; Dignum et al., 2000), and on the probability of being caught by a defender of a norm (Boella and Lesmo, 2001). We consider these cases as very specific and, therefore, inadequate to model different kinds of normative behaviour of autonomous agents.

As a way to overcome these omissions, we have developed a normative framework for agent-based systems that includes a canonical model of norms, a model of normative multi-agent systems and a model of normative autonomous agents. Independent components of this framework have already been presented in different forums (López and Luck, 2003, 2004; López et al., 2002, 2004). The objective of this paper is to present the framework as a whole. The formal model presented in this paper is written in the Z language, which is based on set-theory and first order logic (Spivey, 1992). The organisation of the paper is as follows. First, a formal definition of an autonomous agents is given. After that, an analysis of different properties of norms is provided. This analysis is then used to justify the elements that a general model of a norm must include in order to enable autonomous agents to reason about them. Next, the main properties of systems of autonomous agents that are regulated by norms are discussed and a model is presented. Then, we describe our proposal to enable agents to reason about norms. Finally, our conclusions are provided.

2 Autonomous Agents

The foundations of this work are taken from Luck and d'Inverno's SMART agent framework (d'Inverno and Luck, 2003) whose concept of *motivations* as the driving force that affects the reasoning of agents in satisfying their goals is considered as the underlying argument for agents to voluntarily comply with norms and to voluntarily enter and remain in a society. In the SMART agent framework, an *attribute* represents a perceivable feature of the agent's environment, which can be 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 environmental states, goals, actions and autonomous agents. Details of the remaining elements are not needed, so we simply consider them as given sets.

$$[Attribute, Motivation]$$

$$\begin{aligned}
&EnvState == \mathbb{P}_1\ Attribute\\
&Goal == \mathbb{P}_1\ Attribute\\
&Action == EnvState \rightarrow EnvState
\end{aligned}$$

$$\begin{array}{|l}
\hline
\textit{AutonomousAgent}\\
goals : \mathbb{P}\ Goal; \quad capabilities : \mathbb{P}\ Action;\\
motivations : \mathbb{P}\ Motivation;\\
beliefs : \mathbb{P}_1\ Attribute\\
importance : \mathbb{P}(\mathbb{P}\ Goal \times \mathbb{P}\ Motivation) \rightarrow \mathbb{N}\\
\hline
goals \neq \varnothing; \qquad motivations \neq \varnothing\\
\forall\, x : \mathbb{P}\ Goal, y : \mathbb{P}\ Motivation \bullet\\
\quad (x, y) \in \operatorname{dom} importance \mid\\
\quad\quad x \subseteq goals \wedge y \subseteq motivations\\
\hline
\end{array}$$

In the above schema, an autonomous agent is described by a set of goals that it wants to bring about, a set of capabilities that it is able to perform, a non-empty set of motivations representing its preferences, and a set of beliefs representing its vision about the external world. We also assume that the agent is able to determine the $importance$ of its goals, which depends on its current motivations.

3 Norms

Norms facilitate mechanisms to drive the behaviour of agents, especially in those cases when their behaviour affects other agents. Norms can be characterised by their *prescriptiveness*, *sociality*, and *social pressure*. In other words,

- a norm tells an agent how to behave (*prescriptiveness*);
- in situations where more than one agent is involved (*sociality*);
- and since it is always expected that norms conflict with the personal interest of some agents, socially acceptable mechanisms to force agents to comply with norms are needed (*social pressure*).

By analysing these properties, the essential components of a norm can be identified.

3.1 Norm Components

Norms specify patterns of behaviour for a set of agents. These patterns are sometimes represented as actions to be performed (Axelrod, 1986; Tuomela, 1995), or restrictions to be imposed over an agent's actions (Norman et al., 1998; Shoham and Tennenholtz, 1995). At other times, patterns of behaviour are specified through goals that must either be satisfied or avoided by agents (Conte and Castelfranchi, 1995; Singh, 1999). Now, since actions are performed in order to change the state of an environment, goals are states that agents want to bring about, and restrictions can be seen as goals to be avoided, we argue that by considering goals the other two patterns of behaviour can be easily represented (as shown in (López and Luck, 2003)).

In brief, norms specify things that ought to be done and, consequently, a set of *normative goals* must be included. Sometimes, these normative goals must be directly intended, while at other times their role is to inhibit specific states (as in the case of prohibitions). Norms are always directed at a set of *addressee agents*, which are directly responsible for the satisfaction of the normative goals. Moreover, sometimes to take decisions regarding norms, agents not only consider what must be done but also for whom it must be done. Then, agents that *benefit* from the satisfaction of normative goals may also be included.

In general, norms are not applied all the time, but only in particular circumstances or within a specific *context*. Thus, norms must always specify the situations in which addressee agents must fulfill them. *Exception* states may also be included to represent situations in which addressees cannot be punished when they *have not* complied with norms. Exceptions represent *immunity* states for all addressee agents in a particular situation (Ross, 1968). Now, to ensure that personal interests do not impede the fulfillment of norms, mechanisms either to promote compliance with norms, or to inhibit deviation from them, are needed. Norms may include *rewards* to be given when normative goals become satisfied, or *punishments* to be applied when they are not. Both rewards and punishments are the means for addressee agents to determine what might happen whatever decision they take regarding norms. They are not the responsibility of addressees agents but of other agents already entitled to either reward or punish compliance and non-compliance with norms. Since rewards and punishments represent states to be achieved, it is natural to consider them as goals but, in contrast with normative goals that must be satisfied by addressees, punishments and rewards are satisfied by agents entitled to do so.

In other words, a norm must be considered for fulfillment by an agent when certain environmental states, not included as exception states, hold. Such a norm forces a group of addressee agents to satisfy some normative goals for a (possibly empty) set of beneficiary agents. In addition, agents are aware that rewards may be enjoyed if norms become satisfied, or that punishments that affect their current goals can be applied if not. The formal specification of a norm is given in the $Norm$ schema where all the components of norms described here are included, together with some constraints on them. First, it does not make any sense to have norms specifying nothing, norms directed at nobody, or norms that either never or always become applied. Thus, the first three predicates in the schema state that the set of normative goals, the set of addressee agents, and the context must never be empty. The fourth predicate states that the set of attributes describing both the context and exceptions must be disjoint to avoid inconsistencies in identifying whether a norm must be applied. The final constraint specifies that punishments and rewards are also consistent and, therefore, they must be disjoint.

$$\begin{array}{|l}
\multicolumn{1}{l}{\mathit{Norm}} \\
\mathit{normativegoals} : \mathbb{P}\, \mathit{Goal} \\
\mathit{addressees} : \mathbb{P}\, \mathit{NormativeAgent} \\
\mathit{beneficiaries} : \mathbb{P}\, \mathit{NormativeAgent} \\
\mathit{context} : \mathit{EnvState} \\
\mathit{exceptions} : \mathit{EnvState} \\
\mathit{rewards} : \mathbb{P}\, \mathit{Goal} \\
\mathit{punishments} : \mathbb{P}\, \mathit{Goal} \\
\hline
\mathit{normativegoals} \neq \varnothing \\
\mathit{addressees} \neq \varnothing \\
\mathit{context} \neq \varnothing \\
\mathit{context} \cap \mathit{exceptions} = \varnothing \\
\mathit{rewards} \cap \mathit{punishments} = \varnothing \\
\hline
\end{array}$$

3.2 Considerations

The term *norm* has been used as a synonym for obligations (Boella and Lesmo, 2001; Dignum et al., 2000), prohibitions (Dignum, 1999), social laws (Shoham and Tennenholtz, 1995), and other kinds of rules imposed by societies (or by an authority). The position of our work is quite different. It considers that all these terms can be grouped in a general definition of a norm, because they have the same properties (i.e. prescriptiveness, sociality and social pressure) and they can be represented by the same model. They all represent responsibilities for addressee agents, and create expectations for beneficiaries and other agents. They are also the means to support beneficiaries when they have to claim some compensation in the situations where norms are not fulfilled as expected. Moreover, whatever the kind of norm being considered, its fulfillment may be rewarded, and its violation may be penalised. What makes one norm different from another is the way in which they are created, their persistence, and the components that are obligatory in the norm. Thus, norms might be created by an agent designer as built-in norms, they can be the result of agreements between agents, or they can be elaborated by a complex legal system. Regarding their persistence, norms might be taken into account during different periods of time, such as until an agent dies, as long as an agent stays in a society, or just for a short period of time until its normative goals become satisfied. Finally, some components of a norm might not exist; there are norms that include neither punishments nor rewards, even though they are complied with. Some of these characteristics can be used to provide a *classification* of norms into four main categories: *obligations*, *prohibitions*, *social commitments* and *social codes*. Despite these differences, all types of norms can be reasoned about in similar ways.

Now, to understand the consequences of norms in a particular system, it is necessary to consider norms that are either *fulfilled* or *unfulfilled*. However, since most of the time a norm has a set of agents as addressees, the meaning of fulfilling a norm might depend on the interpretation of analysers of a system. In small groups of agents, it might be easy to consider a norm as fulfilled when every addressee agent has fulfilled the norm; by contrast, in larger societies, a proportion of agents complying with a norm will be enough to consider it as fulfilled. Instead of defining fulfilled norms in general, it is more appropriate to define norms being fulfilled by a particular addressee agent. To do so, the concept of norm instances is introduced as follows. Once a norm is adopted by an agent, a *norm instance* is created, which represents the internalisation of a norm by an agent (Conte and Castelfranchi, 1995). A norm instance is a copy of the original norm that is now used as a *mental attitude* from which new goals for the agent might be inferred. Norms and norm instances are the same concept used for different purposes. Norms are abstract specifications that exist in a society and are known by all agents (Tuomela, 1995), but agents work with *instances* of these norms. Consequently, there must be a separate instance for each addressee of a norm. Due to space constraints, formal definitions and examples of categories of norms, norm instances and fulfilled norms are not provided here but can be found elsewhere (López and Luck, 2003).

3.3 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 (Ross, 1968). 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.

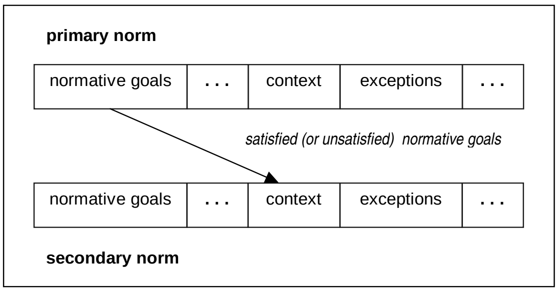


Figure 1: Interlocking Norm Structure

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, where n_1 represents the primary norm and n_2 is the secondary norm.

$$\begin{array}{|l}
\mathit{lockedbynoncompliance}\_ : \mathbb{P}(\mathit{Norm} \times \mathit{Norm}) \\
\hline
\forall\, n_1, n_2 : \mathit{Norm} \bullet \\
\quad \mathit{lockedbynoncompliance}\ (n_1, n_2) \Leftrightarrow \\
\qquad (\exists\, \mathit{ni} : \mathit{NormInstance} \mid \\
\qquad \mathit{isnorminstance}\ (\mathit{ni}, n_1) \bullet \\
\qquad\quad \neg\ \mathit{fulfilled}\ (\mathit{ni}, n_2.\mathit{context}))
\end{array}$$

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.

$$\begin{array}{|l}
\mathit{lockedbycompliance}\_ : \mathbb{P}(\mathit{Norm} \times \mathit{Norm}) \\
\hline
\forall\, n_1, n_2 : \mathit{Norm} \bullet \\
\quad \mathit{lockedbycompliance}\ (n_1, n_2) \Leftrightarrow \\
\qquad (\exists\, \mathit{ni} : \mathit{NormInstance} \mid \\
\qquad \mathit{isnorminstance}\ (\mathit{ni}, n_1) \bullet \\
\qquad\quad \mathit{fulfilled}\ (\mathit{ni}, n_2.\mathit{context}))
\end{array}$$

Having the means to relate norms in this way allows us to model how the normative behaviour of agents that are addressees of a secondary norm is *influenced* by the normative behaviour of addressees of a primary norm.

4 Normative Multi-Agent Systems

Since norms are social concepts, they cannot be studied independently of the systems for which they are created and, consequently, an analysis of the normative aspects of social systems must be provided. Although social systems that are regulated by norms are different from one another, some general characteristics can be identified. They consist of a set of agents that are controlled by the same set of norms ranging from obligations and social commitments to social codes. However, whereas there are static systems in which all norms are defined in advance and agents in the system always comply with them (Boman, 1999; Shoham and Tennenholtz, 1995), a more realistic view of these kinds of systems suggests that when *autonomous* agents are considered, neither can all norms be known in advance (since new conflicts among agents may emerge and, therefore, new norms may be needed), nor can compliance with norms be guaranteed (since agents can decide not to comply). We can say then, that systems regulated by norms must include mechanisms to deal with both the modification of norms and the unpredictable normative behaviour of autonomous agents. So, *normative multi-agent systems* have the following characteristics.

- *Membership*. Agents in a society must be able to deal with norms but, above all, they must recognise themselves as part of the system. This kind of social identification means that agents adopt the society norms and, by doing so, they show their willingness to comply with these norms.
- *Social Pressure*. Effective authority cannot be exerted if penalties or incentives are not applied when norms are either violated or complied with. However, this control must not be an agent's arbitrary decision, and although it is only exerted by some agents, it must be socially accepted.
- *Dynamism*. Normative systems are *dynamic* by nature. New norms are created and obsolete norms are abolished. Compliance or non-compliance with norms may activate other norms and, therefore, force other agents to act. Agents can either join or leave the system. The normative behaviour of agent members might be unexpected, and it may influence the behaviour of other agents.

Given these characteristics, we argue that multi-agent systems must include mechanisms to defend norms, to allow their modification, and to identify authorities. Moreover, their members must be agents able to deal with norms. Each one of these concepts is discussed in detail and formalised in (López and Luck, 2004), here, we present just a summary of them.

4.1 Normative Agents

The effectiveness of every structure of control relies on the capabilities of its members to recognise and follow its norms. However, given that agents are autonomous, the fulfillment of norms can never be taken for granted (López et al., 2002). A *normative agent* is an agent whose behaviour is partly shaped by norms. They are able to deal with norms because they can represent, adopt, and comply with them. However, for autonomous agents, decisions to adopt or comply with norms are made on the basis of their own goals and motivations. That is, autonomous agents are not only able to *act on* norms but also they are able to *reason about* them. In what follows, all normative agents are considered as autonomous agents that have adopted some norms (*norms*) and, has decided which norms to comply with (*intended norms*) and which norms to reject (*rejected norms*). Although their normative behaviour is described in the next section, their representation is given now in the schema below.

$$\begin{array}{|l}
\hline
\mathit{NormativeAgent} \\
\mathit{AutonomousAgent} \\
\mathit{norms}, \mathit{intented}, \mathit{rejected} : \mathbb{P}\, \mathit{Norm} \\
\hline
\mathit{intented} \subseteq \mathit{norms} \\
\mathit{rejected} \subseteq \mathit{norms} \\
\hline
\end{array}$$

4.2 Enforcement and Reward Norms

Particularly interesting for this work are the norms triggered in order to punish offenders of other norms. We call them *enforcement norms* and their addressees are the *defenders* of a norm. These norms represent exerted social pressure because they specify not only who must apply the punishments, but also under which circumstances these punishments must be applied (Ross, 1968). That is, once the violation of a norm becomes identified by defenders, their duty is to start a process in which offender agents can be punished. 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 refuses to pay.

As can be seen, norms that enforce other norms are a special case of interlocking norms because besides being interlocked by non-compliance, the normative goals of the secondary norm must include every punishment of the primary norm. Figure 2 shows how the structures of both norms are related. By modelling enforcement norms in this way, we cause an offender's punishments to be consistent with a defender's responsibilities. Addressees of an *enforced* norm (i.e. the primary norm) know what could happen if the norm is not complied with, and addressees of an *enforcement* norm (i.e. the secondary norm) know what must be done in order to punish the offenders of another norm. Enforcement norms allow the authority of defenders to be clearly constrained.

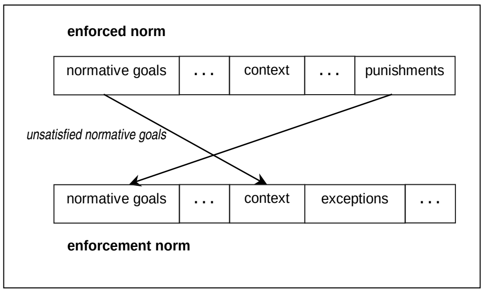


Figure 2: Enforcement Norm Structure

Formally, the relationship between a norm directed to control the behaviour of some agents and a norm directed at punishing the offenders of such a norm can be defined as follows. A norm *enforces* another norm if the first norm is activated when the second is violated, and all punishments associated with the violated norm are part of the normative goals of the first. Every norm satisfying this property is known as an *enforcement* norm.

$$enforces\_ : \mathbb{P}(Norm \times Norm)$$

$$\forall\, n_1, n_2 : Norm \bullet enforces\ (n_1, n_2) \Leftrightarrow$$
$$lockedbynoncompliance\ (n_2, n_1) \wedge$$
$$n_2.punishments \subseteq n_1.normativegoals$$

So far we have described some interlocking norms in terms of punishments because these are one of the more commonly used mechanisms to enforce compliance with norms. However, a similar analysis can be applied to interlocking norms corresponding to the process of rewarding members doing their duties. These norms must be interlocked by compliance and all the rewards included in the primary norm (rewarded norm) must be included in the normative goals of the secondary norm (reward norm). The relation between these norms is shown in Figure 3.

Formally, we say that a norm *encourages* compliance with another norm if the first norm is activated when the second norm becomes fulfilled, and the rewards associated with the fulfilled norm are part of the normative goals of the first norm. Every norm satisfying this property is known as a *reward* norm.

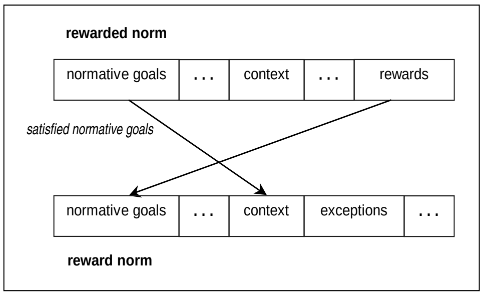


Figure 3: Reward Norm Structure

$$rewardnorm\_ : \mathbb{P}(Norm \times Norm)$$

$$\forall\, n_1, n_2 : Norm \bullet rewardnorm\ (n_1, n_2) \Leftrightarrow$$
$$lockedbycompliance\ (n_2, n_1) \wedge$$
$$n_2.rewards \subseteq n_1.normativegoals$$

It is important to mention that this way of representing enforcement and reward norms can create an infinite chain of norms because we would also have to define norms to apply when authorities or defenders do not comply with their obligations, either to punish those agents breaking rules or to reward those agents that fulfill their responsibilities (Ross, 1968). The decision of when to stop this interlocking of norms is left to the creator of norms. If a system requires it, the 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.

Both enforcement and reward norms acquire particular relevance in systems regulated by norms because the abilities to punish and reward must be restricted for use only by competent authorities (addressees of enforcement and reward norms). Otherwise, offenders might be punished twice or more if many agents take this as their responsibility. It could also be the case that selfish agents demand unjust punishments or that selfish offenders reject being punished. That is, conflicts of interest might emerge in a society if such responsibilities are given either to no one or to anyone. Only through enforcement and reward norms can agents become entitled to punish or reward other agents.

4.3 Legislation Norms

Norms are introduced into a society as a means to achieve social order. Some 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 all commitments can be anticipated. Consequently, there must exist the possibility of creating new norms (to solve unexpected and recurrent conflicts among agents), modifying existing ones (to increase their effectiveness), or even abolishing those that become obsolete. As above, these capabilities must be restricted to avoid conflicts of interest. That is, norms stating when actions to legislate are permitted must exist in a normative multi-agent system (Jones and Sergot, 1996). Formally, we say that a norm is a *legislation* norm if actions to issue and to abolish norms are permitted by this norm in the current environment. These constraints are specified below.

$$\begin{array}{|l}
legislate\_ : \mathbb{P}(Norm \times EnvState) \\
\hline
\forall\, n : Norm;\ env : EnvState \bullet \\
legislate\ (n, env) \Leftrightarrow \\
\quad (\exists\, issuingnorms, abolishnorms : Action \bullet \\
\qquad permitted\ (issuingnorms, n, env) \lor \\
\qquad permitted\ (abolishnorms, n, env))
\end{array}$$

4.4 Normative Multi-Agent Systems Model

A normative multi-agent system is formally represented in the $NormativeMAS$ schema. It comprises a set of normative agent members (i.e. agents able to reason about norms) and a set of general norms that govern the behaviour of these agents ($generalnorms$). Norms issued to allow the creation and abolition of norms ($legislationnorms$) are also included. There are also norms dedicated to enforcing other norms ($enforcenorms$) and norms directed to encouraging compliance with norms through rewards ($rewardnorms$). Legislation, enforcement and reward norms are better discussed in (López and Luck, 2004). The current state of the environment is represented by the variable $environment$. Constraints over these components are imposed as follows. Although it is possible that agents do not know all the norms in the system, it is always expected that they at least adopt some norms, represented by the first predicate. The second predicate makes explicit that addressees of norms must be members of the system. Thus, 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 nonexistent agents. 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, in which case no one in the society is legally entitled to punish an agent that does not fulfill such a norm.

$$\begin{array}{|l}
\multicolumn{1}{l}{NormativeMAS} \\
members : \mathbb{P}\, NormativeAgent \\
generalnorms, legislationnorms : \mathbb{P}\, Norm \\
enforcenorms, rewardnorms : \mathbb{P}\, Norm \\
environment : EnvState \\
\hline
\forall\, ag : members \bullet \\
\quad ag.norms \cap generalnorms \neq \varnothing \\
\forall\, sn : generalnorms \bullet \\
\quad sn.addressees \subseteq members \\
\forall\, en : enforcenorms \bullet \\
\quad (\exists\, n : generalnorms \bullet enforces\ (en, n)) \\
\forall\, rn : rewardnorms \bullet \\
\quad (\exists\, n : generalnorms \bullet rewardnorm\ (rn, n)) \\
\forall\, ln : legislationnorms \bullet \\
\quad legislate\ (ln, environment)
\end{array}$$

4.5 Normative Roles

Defining normative multi-agent systems in this way allows the identification of the *authorities* of the system as formalised in the $AuthoritiesNMAS$ schema. The set of agents that are entitled to create, modify, or abolish norms is called *legislators*. No other members of the society are endowed with this authority, and generally they are either elected or imposed by other agents. *Defender* agents are directly responsible for the application of punishments when norms are violated. That is, their main responsibility is to monitor compliance with norms in order to detect transgressions. Moreover, they can also warn agents by advertising the bad consequences of being rebellious. By contrast, *promoter* agents are those whose responsibilities include rewarding compliant addressees. These agents also monitor compliance with norms in order to determine when rewards must be given, and instead of *enforcing* compliance with norms, they simply *encourage* it.

$$\begin{array}{|l}
\multicolumn{1}{l}{AuthoritiesNMAS} \\
NormativeMAS \\
legislators : \mathbb{P}\, NormativeAgent \\
defenders : \mathbb{P}\, NormativeAgent \\
promoters : \mathbb{P}\, NormativeAgent \\
\hline
\forall\, lg : legislators \bullet (\exists\, l : legislationnorms \bullet \\
\quad lg \in l.addressees) \\
\forall\, df : defenders \bullet (\exists\, e : enforcenorms \bullet \\
\quad df \in e.addressees) \\
\forall\, pm : promoters \bullet (\exists\, r : rewardnorms \bullet \\
\quad pm \in r.addressees)
\end{array}$$

5 Autonomous Normative Reasoning

Whereas agents that always comply with norms are important for the design of societies in which total control is needed (Boman, 1999; Shoham and Tennenholtz, 1995), agents that can decide on the basis of their own goals and motivations whether to comply with them are important for the design of dynamic systems in which agents act on behalf of different users and, while satisfying their own goals, are able to join a society and cooperate with other agents. Autonomous norm reasoning is important to address those situations in which an agent's goals conflict with the norms that control its behaviour inside a society. Agents that deliberate about norms are also needed in systems in which unforseen events might occur, and in those situations in which agents are faced with conflicting norms, and they have to choose between them. It should be clear that violation of norms is, sometimes, justified. To describe *normative reasoning*, therefore, we have to explain not only what might motivate an agent to adopt, dismiss or complying with a norm, but also the way in which this decision affects its goals. In consequence we propose three different processes: one for agents to decide whether to adopt a norm (*the norm adoption process*), another to decide whether to comply with a norm (*the norm deliberation process*), and the other to update the goals, and therefore the intentions of agents accordingly (*the norm compliance process*). All these processes must take into account not only the goals and motivations of agents, but also the mechanisms of the society to avoid violation of norms such as rewards and punishments. Thus, agents consider the so called *social pressure of norms* before making any decision.

5.1 The Norm Adoption Process

The *norm adoption* process can be better defined as the process through which agents recognise their responsibilities towards other agents by internalising the norms that specify these responsibilities. Thus, agents adopt the norms of a society either once they have decided to join it or in the case a new norm is issued while they are still there. For autonomous agents to join and stay in a society the *social satisfaction condition* must hold (López et al., 2004). An agent considers this condition as satisfied if, although some of its goals become hindered by its *responsibilities*, its important goals can still be satisfied. Thus, we consider that the following conditions must be satisfied for agents to adopt a norm: the agent must recognise itself as an addressee of the norm; the norm must not already be adopted; the norm must have been issued by a recognised authority; and the agent must have reasons to stay in the society. Notice that to adopt a norm as an end, only the first three conditions are needed, whereas the last condition is an indicator that the decision to adopt a norm is made in an autonomous way. Due to space constraints, the $NormAdoption$ schema only formalises the first three conditions but details of the fourth condition can be found elsewhere (López et al., 2004).

$$\begin{array}{|l}
\multicolumn{1}{l}{\textit{NormAdoption}} \\
\hline
\Delta \textit{NormativeAgent} \\
\textit{new}? : \textit{Norm} \\
\textit{issuer}?, \textit{self} : \textit{NormativeAgent} \\
\textit{authorities} : \mathbb{P}\, \textit{NormativeAgent} \\
\textit{issuedby} : \mathbb{P}(\textit{Norm} \times \textit{NormativeAgent}) \\
\hline
\textit{self} \in \textit{new}?.\textit{addressees} \\
\textit{new}? \notin \textit{norms} \\
(\textit{new}?,\ \textit{issuer}?) \in \textit{issuedby} \Leftrightarrow \\
\qquad \textit{issuer}? \in \textit{authorities} \\
\textit{norms}' = \textit{norms} \cup \{\textit{new}?\} \\
\hline
\end{array}$$

5.2 The Norm Deliberation Process

To comply with the norm, agents assess two things: the goals that might be hindered by satisfying the normative goals, and the goals that might benefit from the associated rewards. By contrast, to reject a norm, agents evaluate the damaging effects of punishments (i.e. the goals hindered due to the satisfaction of the goals associated with punishments.) Since the satisfaction of some of their goals might be prevented in both cases, agents use the *importance* of their goals to make these decisions. This, to deliberate about a norm, an agents pursues the following steps.

- A set of *active* norms is selected from the set of adopted norms (norm instances). Active norms are those that agents believe must be complied with in the current state, which is not an exception state (i.e. those norms for which the context matches the beliefs of the agent).
- The agent divides active norms into *non-conflicticting* and *conflicting* norms. An active norm is *non-conflicting* if its compliance does not cause any conflict with one of the agent's current goals. Thus, no goals of the addressee agent are hindered by satisfying the normative goals of the norm. By contrast, an active norm is *conflicting* if its fulfillment hinders any of the agent's goals.
- For each one of these sets of norms, the agent must decide which one to comply with. Details of different ways to select the norms to be

intended or rejected are given in (López et al., 2002). After norm deliberation, the set of intended norms consists of those conflicting and non-conflicting norms that are accepted to be complied with by the agent, and the set of rejected norms consists of all conflicting and non-conflicting norms that are rejected by the agent.

The state of an agent that has selected the norms it is keen to fulfill is formally represented in the $\mathit{NormAgentState}$ schema. This represents a normative agent with a variable representing the sets of active norms at a particular point of time. The $\mathit{conflicting}$ predicate holds for a norm if and only if its normative goals conflict (hinder) with any of the agent's current goals. The next three predicates state that active norms are the subset of adopted norms that the agent believes must be complied with in the current state and that, the set of active norms has already been assessed and divided into norms to intend and norms to reject. The state of an agent is consistent in that its current goals do not conflict with the intended norms and, consequently, no normative goal must be in conflict with current goals. Moreover, since rewards benefit the achievement of some goals, so that agents do not have to work on their satisfaction because someone else does, these goals must not be part of the goals of an agent. The final predicate states that punishments must be accepted and, therefore, none of the goals of an agent must hinder them.

$$\begin{array}{|l}
\multicolumn{1}{l}{\mathit{NormAgentState}} \\
\hline
\mathit{NormativeAgent} \\
\mathit{activenorms}, \mathit{conflicting}\ \_ : \mathbb{P}\ \mathit{Norm} \\
\hline
\forall\ n : \mathit{activenorms} \bullet \mathit{conflicting}\ n \Leftrightarrow \\
\qquad \mathit{hinder}(\mathit{goals}, n.\mathit{ngoals}) \neq \varnothing \\
\mathit{activenorms} \subseteq \mathit{norms} \\
\forall\ \mathit{an} : \mathit{activenorms} \bullet \\
\qquad \mathit{logcon}\ (\mathit{beliefs}, \mathit{an}.\mathit{context}) \\
\mathit{activenorms} = \mathit{intended} \cup \mathit{rejected} \\
\mathit{hinder}(\mathit{goals}, \mathit{normgoals}\ \mathit{intended}) = \varnothing \\
\mathit{benefit}(\mathit{goals}, \mathit{rewardgoals}\ \mathit{intended}) \\
\qquad \cap \mathit{goals} = \varnothing \\
\mathit{hinder}(\mathit{goals}, \mathit{punishgoals}\ \mathit{rejected}) = \varnothing \\
\hline
\end{array}$$

For a norm to be intended, some constraints must be fulfilled. First, the agent must be an addressee of the norm. Then, the norm must be an adopted and currently active norm, and it must not be already intended. In addition, the agent must believe that it is not in an *exception* state and, therefore, it must comply with the norm. Formally, the process to accept a single norm as input ($\mathit{new}?$) to be complied with is specified in the $\mathit{NormIntend}$ schema. The first five predicates represent the constraints on the agent and the norm as described above. The sixth predicate represents the addition of the accepted norm to the set of intended norms and the final predicate represents the set of rejected norms remains the same.

$$\begin{array}{|l}
\multicolumn{1}{l}{\mathit{NormIntend}} \\
\hline
\mathit{new}? : \mathit{Norm} \\
\Delta \mathit{NormAgentState} \\
\hline
\mathit{self} \in \mathit{new}?.\mathit{addressees} \\
\mathit{new}? \in \mathit{norms} \\
\mathit{new}? \in \mathit{activenorms} \\
\mathit{new}? \notin \mathit{intended} \\
\neg\ \mathit{logcon}(\mathit{beliefs}, \mathit{new}?.\mathit{exceptions}) \\
\mathit{intended}' = \mathit{intended} \cup \{\mathit{new}?\} \\
\mathit{rejected}' = \mathit{rejected} \\
\hline
\end{array}$$

The process to reject a norm ($\mathit{NormReject}$) can be defined similarly. Now, there are different ways to select the norms to be intended or rejected as explained in (López et al., 2002). Here, we describe what is called a *pressured* strategy where an agent fulfills a norm only in the case that one of its goals is threatened by punishments. That is, agents are *pressured* to obey norms through the application of punishments that might hinder some of their important goals. In this situation, the agent faces four different cases.

1. The norm is a non-conflicting norm and some goals are hindered by its punishments.
2. The norm is a non-conflicting norm and there are no goals hindered by its punishments.
3. The norm is a conflicting norm and the goals hindered by its normative goals are less important than the goals hindered by its punishments.
4. The norm is a conflicting norm and the goals hindered by its normative goals are more important than the goals hindered by its punishments.

The first case represents the situation in which, by complying with a norm, an agent does not put at risk any of its goals (because the norm is non-conflicting), but if the agent decides not to fulfill it, some of its goals could be unsatisfied due to punishments. Consequently, fulfilling a norm is the best decision for this kind of agent. To formalise this, we use the $\mathit{NormIntend}$ operation schema to accept complying with the norm, and we add two predicates to specify that this strategy is applied to non-conflicting norms whose punishments hinder some goals.

$$\begin{array}{|l}
\multicolumn{1}{l}{\mathit{PressuredNCComply}} \\
\hline
\mathit{NormIntend} \\
\hline
\neg\ \mathit{conflicting}\ \mathit{new}? \\
\mathit{hinder}(\mathit{goals}, \mathit{new}?.\mathit{punishments}) \neq \varnothing \\
\hline
\end{array}$$

In the second case, by contrast, since punishments do not affect an agent's goals, it does not make any sense to comply with the norm, so it must be rejected. Formally, the $\mathit{NormReject}$ operation schema is used when the norm is non-conflicting (first predicate) and its associated punishments do not hinder any existing goals (second predicate).

$$
\begin{array}{|l}
\mathit{PressuredNCReject} \\
\hline
\mathit{NormReject} \\
\hline
\neg\ \mathit{conflicting}\ \mathit{new?} \\
\mathit{hinder}(\mathit{goals}, \mathit{new?.punishments}) = \varnothing \\
\hline
\end{array}
$$

According to our definition, a conflicting norm is a norm whose normative goals hinder an agent's goals. In this situation, agents comply with the norm at the expense of existing goals only if what they can lose through punishments is more important than what they can lose by complying with the norm. Formally, a conflicting norm is intended if the goals that could be hindered by punishments (hps) are more important than the set of existing goals hindered by normative goals (hngs). This is represented in the $\mathit{PressuredCComply}$ schema where the $\mathit{importance}$ function uses the motivations associated with the set of goals to find the importance of goals.

$$
\begin{array}{|l}
\mathit{PressuredCComply} \\
\hline
\mathit{NormIntend} \\
\hline
\mathit{conflicting}\ \mathit{new?} \\
\mathbf{let}\ \mathit{hps} == \mathit{hinder}(\mathit{goals}, \\
\quad \mathit{new?.punishments}) \bullet \\
\mathbf{let}\ \mathit{hngs} == \mathit{hinder}(\mathit{goals}, \mathit{new?.ngoals}) \bullet \\
\quad \mathit{importance}\ (\mathit{motivations}, \mathit{hps}) > \\
\qquad \mathit{importance}\ (\mathit{motivations}, \mathit{hngs}) \\
\hline
\end{array}
$$

However, if the goals hindered by normative goals are more important than the goals hindered by punishment, agents prefer to face such punishments for the sake of their important goals and, therefore, the norm is rejected. Formally, a conflicting norm is rejected by using the $\mathit{NormReject}$ operation schema if the goals hindered by its punishments (hps) are less important than the goals hindered by its normative goals (hngs).

$$
\begin{array}{|l}
\mathit{PressuredCReject} \\
\hline
\mathit{NormReject} \\
\hline
\mathit{conflicting}\ \mathit{new?} \\
\mathbf{let}\ \mathit{hps} == \mathit{hinder}(\mathit{goals}, \mathit{new?.punishments}) \bullet \\
\mathbf{let}\ \mathit{hngs} == \mathit{hinder}(\mathit{goals}, \mathit{new?.ngoals}) \bullet \\
\quad \mathit{importance}\ (\mathit{motivations}, \mathit{hps}) \leq \\
\qquad \mathit{importance}\ (\mathit{motivations}, \mathit{hngs}) \\
\hline
\end{array}
$$

All these cases are illustrated in Figure 4

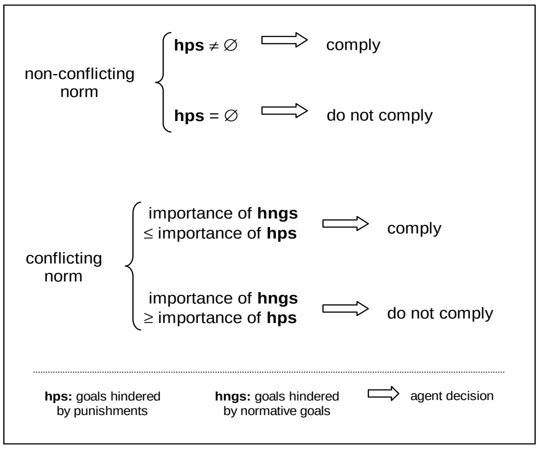

non-conflicting norm
hps ≠ ∅
comply
hps = ∅
do not comply
conflicting norm
importance of hngs ≤ importance of hps
comply
importance of hngs ≥ importance of hps
do not comply
hps: goals hindered by punishments
hngs: goals hindered by normative goals
agent decision

Figure 4: Pressured Norm Compliance

5.3 The Norm Compliance Process

Once agents take a decision about which norms to fulfill, a process of *norm compliance* must be started in order to update an agent's goals in accordance with the decisions it has made. An agent's goals are affected in different ways, depending on whether the norm is intended or rejected. The cases can be listed as follows.

- All normative goals of an intended norm must be added to the set of goals because the agent has decided to comply with it.
- Some goals are hindered by the normative goals of an intended norm. These goals can no longer be achieved because the agent prefers to comply with the norm and, consequently, this set of goals must be removed from the agent's goals.
- Some goals benefit from the rewards of an intended norm. Rewards contribute to the satisfaction of these goals without the agent having to make any extra effort. As a result, those goals that benefit from rewards must no longer be considered by the agent to be satisfied, and must be removed from the set of goals.
- Rejected norms only affect the set of goals hindered by the associated punishments. This set of goals must be removed; this is the way in which normative agents accept the consequences of their decisions.

To make the model simple, we assume that punishments are always applied, and rewards are always given, though the possibility exists that agents never become either punished or rewarded. In addition, note that the set of goals hindered by normative goals can be empty if the norm being considered is a non-conflicting norm, and goals hindered by punishments or goals that benefit from rewards can be empty if a norm does not include any of them. After norm compliance, the goals are updated and, consequently, the intentions of agents might change. The process to comply with the norms an agent has decided to fulfill is specified in the $NormComply$ schema. Through this process, the set of goals is updated according to our discussion above.

$$\begin{array}{|l}
\hline NormComply \\
\Delta NormAgentState \\
\hline
\textbf{let } ngs == \bigcup\{gs : \mathbb{P}\ Goal \mid \\
\quad (\exists\, n : intended \bullet gs = n.ngoals)\} \bullet \\
\textbf{let } hngs == \bigcup\{gs : \mathbb{P}\ Goal \mid (\exists\, n : intended \bullet \\
\quad gs = hinder\ (goals, n.ngoals))\} \bullet \\
\textbf{let } brs == \bigcup\{gs : \mathbb{P}\ Goal \mid (\exists\, n : intended \bullet \\
\quad gs = benefit(goals, n.rewards))\} \bullet \\
\textbf{let } hps == \bigcup\{gs : \mathbb{P}\ Goal \mid (\exists\, n : rejected \bullet \\
\quad gs = hinder\ (goals, n.punishments))\} \bullet \\
(\ \ goals' = (goals \cup ngs) \setminus \\
\qquad\qquad (hngs \cup brs \cup hps)) \\
\hline
\end{array}$$

6 Conclusions

In this paper, we have presented a normative framework which, besides providing the means to computationally represent many normative concepts, can be used to give a better understanding of norms and normative agent behaviour. The framework explains not only the role that norms play in a society but also the elements that constitute a norm and that, in turn, can be used by agents when decisions concerning norms must be taken. In contrast to other proposals, our normative framework has been built upon the idea of *autonomy* of agents. That is, it is intended to be used by agents that reason about why norms must be adopted, and why an adopted norm must be complied with. Our framework consists of three main components: a canonical model of norms, a model of normative multi-agent systems and a model of normative autonomous agents.

The model of norms differs from others (Boman, 1999; Shoham and Tennenholtz, 1995; Tuomela, 1995) in the way in which patterns of behaviour are prescribed. To describe the pattern of behaviour prescribed by a norm, other models use actions, so that agents are told what exactly they must do. By contrast, we use normative goals, which is an idea more compatible with autonomous agents whose behaviour is driven by goals. Agents can choose the way to satisfy the normative goals, instead of being told exactly how it must be done. Our work also emphasises that all norms can be represented by using similar components, and that they are analysed by agents in similar ways. However, what makes one norm different from another is the way in which norms are created, how long they are valid, and the reasons agents have to adopt them. These factors enable norms to be divided into categories such as obligations and prohibitions, social commitments and social codes.

A collateral result of our work is the proposed model for interlocking norms. These relations between norms have already been mentioned in several papers, especially from philosophical and legal perspectives (Ross, 1968), but no ways to model them have been provided. Dignum's concept of authorisations (Dignum, 1999) attempts to describe norms activated when others are not fulfilled; however, his idea and models are incomplete. We claim that this form of representing connections between norms can be used not only to represent enforcement and reward norms, but also to represent things as complex as contracts and deals among agents.

In contrast to current models of systems regulated by norms (Balzer and Tuomela, 2001; Dignum and Dignum, 2001; Esteva et al., 2001; Shoham and Tennenholtz, 1995) in which no distinction among norms is made, our work emphasises that besides the general norms of the system, at least three kinds of norms are needed, namely norms to legislate, to punish, and to reward other agents. By making this differentiation, agents are able to determine when an issued norm is valid, when an entitled agent can apply a punishment, and who is responsible for giving rewards. In addition, order is imposed on agents responsible for the normative behaviour of other agents, because their authority is defined by the norms that entitle them to exert social pressure. Roles for *legislators*, *defenders*, and *promoters* of norms become easily identified as a consequence of the different kinds of norms considered. Thus, in this framework, the authority of agents is always supported and constrained by norms.

Acknowledgements

The first author acknowledges funding from the Faculty Enhancement Program (PROMEP) of the Mexican Ministry of Public Education (SEP), project No PROMEP/103.5/04/767.

References

R. Axelrod. An evolutionary approach to norms. *The American Political Science Review*, 80(4):1095–1111, 1986.

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, 2001.

M. Barbuceanu, T. Gray, and S. Mankovski. The role of obligations in multiagent coordination. *Applied Artificial Intelligence*, 13(1/2):11–38, 1999.

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, 2001.

M. Boman. Norms in artificial decision making. *Artificial Intelligence and Law*, 7(1):17–35, 1999.

C. Castelfranchi, F. Dignum, C. Jonker, and J. Treur. Deliberative normative agents: Principles and architecture. In N. Jennings and Y. Lesperance, editors, *Intelligent Agents VI*, LNAI 1757, pages 206–220. Springer, 2000.

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

R. Conte, C. Castelfranchi, and F. Dignum. Autonomous norm-acceptance. In J. Müller, M. Singh, and A. Rao, editors, *Intelligent Agents V*, LNAI 1555, pages 319–333. Springer, 1999a.

R. Conte and Ch. Dellarocas. Social order in info societies: An old challenge for innovation. In C. Dellarocas and R Conte, editors, *Social Order in Multi-Agent Systems*, pages 1–15. Kluwer Academic, 2001.

R. Conte, R. Falcone, and G. Sartor. Agents and norms: How to fill the gap? *Artificial Intelligence and Law*, 7(1):1–15, 1999.

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, 2001.

F. Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1):69–79, 1999.

F. Dignum, D. Morley, E. Sonenberg, and L. Cavendon. Towards socially sophisticated BDI agents. In Edmund H. Durfee, editor, *The Fourth International Conference on Multi-Agent Systems*, pages 111–118. IEEE Computer Society, 2000.

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.

M. d'Inverno and M. Luck. *Understanding Agent Systems*. Springer-Verlag, second edition, 2003.

M. Esteva, J. Padget, and C. Sierra. Formalizing a language for institutions and norms. In J. Meyer and M. Tambe, editors, *Intelligent Agents VIII*, LNAI 2333, pages 348–366. Springer, 2001.

A. Jones and M. Sergot. A formal characterisation of institutionalised power. *Logic Journal of the IGPL*, 4(3):429–445, 1996.

F. López y López and M. Luck. Modelling norms for autonomous agents. In E. Chávez, J. Favela, M. Mejía, and A. Oliart, editors, *The Fourth Mexican Conference on Computer Science*, pages 238–245. IEEE Computer Society, 2003.

F. López y López and M. Luck. A model of normative multi-agent systems and dynamic relationships. In G. Lindemann, D. Moldt, and M. Paolucci, editors, *Regulated Agent-Based Social Systems*, LNAI 2934, pages 259–280. Springer-Verlag, 2004.

F. López y López, M. Luck, and M. d'Inverno. Constraining autonomy through norms. In C. Castelfranchi and W.L. Johnson, editors, *The First International Joint Conference on Autonomous Agents and Multi Agent Systems AAMAS'02*, pages 674–681. ACM Press, 2002.

F. López y López, M. Luck, and M. d'Inverno. Normative agent reasoning in dynamic societies. In N. Jennings, C. Sierra, L. Sonenberg, and L Tambe, editors, *The Third International Joint Conference on Autonomous Agents and Multi Agent Systems AAMAS'04*, pages 730–737. ACM Press, 2004.

M. Luck, P. McBurney, and C. Preist. *Agent Technology: Enabling Next Generation Computing (A Roadmap for Agent Based Computing)*. AgentLink, 2003.

T. Norman, C. Sierra, and N. Jennings. Rights and commitments in multi-agent agreements. In Yves Demazeau, editor, *The Third International Conference on Multi-Agent Systems (ICMAS-98)*, pages 222–229. IEEE Computer Society Press, 1998.

A. Ross. *Directives and Norms*. Routledge and Kegan Paul Ltd., 1968.

Y. Shoham and M. Tennenholtz. On social laws for artificial agent societies: Off-line design. *Artificial Intelligence*, 73(1-2):231–252, 1995.

M. Singh. An ontology for commitments in multiagent systems: Toward a unification of normative concepts. *Artificial Intelligence and Law*, 7(1):97–113, 1999.

J. M. Spivey. *The Z Notation: A Reference Manual*. Prentice-Hall, 1992.

R Tuomela. *The Importance of Us: A Philosophical Study of Basic Social Norms*. Stanford University Press, 1995.

Beyond BDI? Brandomian commitments for multi-agent communication

Rodger Kibble*
*Department of Computing
Goldsmiths College
University of London
London SE14 6NW
r.kibble@gold.ac.uk

Abstract

This paper rehearses some arguments in favour of a normative, commitment based semantics for dialogue acts and multi-agent communication, as opposed to more familiar mentalistic accounts based on notions of belief and intention. The main focus of the paper is on identifying an appropriate notion of *propositional commitment*. A case is made for adopting Brandom's framework of normative pragmatics, modelling dialogue states as *deontic scoreboards* which keep track of commitments and entitlements that speakers acknowledge and hearers attribute to other interlocutors.

1 Background

This paper is part of an ongoing attempt (cf Kibble (2004, 2005)) to pull together some parallel strands in argumentation theory, multi-agent systems, and philosophy of language in order to elucidate a notion of linguistic commitment that can find applications in computational semantics. The long-term goal is to develop a framework which can underpin both natural dialogue modelling and artificial multi-agent communication. Consequently, I try to avoid postulating properties that cannot be safely applied to software agents, in particular *belief* and *intention*.

1.1 Beyond belief

The speech act theories of Austin (1962) and Searle (1965) have had unanticipated computational applications, providing the theoretical underpinning for standard semantics for agent communication languages (ACLs). The essence of speech act theory, following Grice's account of "non-natural meaning" (Grice, 1957), is the assumption that agents typically produce utterances with the *intention* of bringing about some change in the *beliefs* of a hearer, and that the hearer's *recognition* of this intention is crucial to the success of the speech act. An early instance of the take-up of this work in AI was the demonstration by Cohen and Perrault (1979) that Searle's systematic analysis of speech acts such as promising, requesting, asserting in terms of preconditions and outcomes, and the beliefs and intentions of participants, could naturally be formalised in terms of planning operators. This has led (via e.g., Cohen and Levesque (1990)) to important developments in the field of multi-agent and human-computer communication, the most prominent being the adoption of Searlean terminology to specify message types in agent communication languages (cf Wooldridge (2000a)). One motivation for basing ACL specifications on speech act theory is to hold out the hope that artificial agents will eventually converse as freely with human clients as with each other. However, agent design in terms of notions such as *belief* and *intention* faces the software engineering problem that it is not generally possible to identify data structures corresponding to beliefs and intentions in heterogenous agents Wooldridge (2000b), let alone a "theory of mind" enabling agents to reason about other agents' beliefs. This problem has been addressed by proposing alternative semantics based on intersubjectively observable notions of *commitments* (Singh, 1999), thus extending the notion of commitment from the sphere of social interaction to that of communication.

In philosophy of language the notion of propositional commitment goes back at least to Hamblin (1970). Singh's proposal is in some ways anticipated by moves in this field to eliminate or at least downgrade mentalistic notions in favour of social constructs (Brandom, 1983, 1994; Habermas, 1984a,b).

Belief has been a problem for philosophy of language for a number of reasons, including:

- Logical omniscience: if someone is held to believe ϕ, should we hold them to believe all logical consequences of ϕ, including all theorems? E.g. from "John believes whales are mammals" does it follow that "John believes whales are mammals or dolphins are telepathic", etc?

- Inconsistency: suppose John believes two incompatible propositions, do we further suppose that he believes any arbitrary proposition, following the standard deductive rule *ex falso sequitur quodlibet*?

- The notion of *mutual belief* threatens an infinite regress: "John believe that Mary believes that John believes... that ϕ".

- The physical state of an agent underdetermines the beliefs that can be attributed to it; that is, belief is not a "narrow" property. It is possible in principle, if unlikely in practice, that two computers could have identical memory contents and go through the same sequence of operations yet be performing different computations, according to how the user interprets their inputs and outputs.

One way to finesse some of these conundra may be to distinguish between what an agent consciously believes, as a psychological state, and what they have *reason to believe* (cf (Lewis, 1969; Sugden, 2003)). So although an agent may have good reason to believe any logical consequences of their conscious beliefs, they may well not be aware of these consequences unless and until someone points them out; which may involve intricate proofs and argumentation. Now, this approach involves an imprecise use of the word *reason*: when we say someone has "reason to believe" something, this could mean either that it would be reasonable (not absurd) for them to hold this belief, or that there are reasons why they *should* hold it. Following Brandom (1994); Lance (1995) let us call the former sense *permissive* and the latter *committive*. To take a hackneyed example: if I believe that President Nixon was both a Republican and a Quaker, and that all Republicans are bellicose while all Quakers are pacifists, I seem thereby compelled to adopt the beliefs that Nixon was a pacifist, and that he was a warmonger. On the other hand, it would not be *reasonable* for me to believe that he held both these positions: I should reconsider my beliefs in the light of his actual record, and perhaps conclude that not all Quakers are pacifists, or that Nixon was not a "true" Quaker. In Brandom's terms, we can be *committed* to incompatible beliefs but not *entitled* to them.

Hamblin's notion of *propositional commitment* has been further developed, apparently independently, in contributions to argumentation theory by Douglas Walton and associates (Walton, 1993; Walton and Krabbe, 1995), and in Robert Brandom's normative pragmatics (Brandom, 1983, 1994, 2000). Commitment is understood in the sense of the following remark:

> ...to assert a proposition may amount to becoming committed to subsequently defending the proposition, if one is challenged to do so by another speaker in dialogue.
>
> (Walton and Krabbe, 1995, p. 31)

It is important to distinguish between psychological notions of commitment as a state of mind, a kind of persistent intention, and socially-oriented notions of commitment as an obligation towards other agents, which may be claimed by an agent and/or attributed by members of the agent society. In common with the works cited above this paper will follow the latter interpretation.

Brandom's analysis is particularly radical in that he argues that talk about propositional ("doxastic") commitments can effectively supplant talk about beliefs and can finesse certain long-standing issues in belief modelling (Brandom, 1994, p. 196).

1. By distinguishing between commitments which an agent *acknowledges* and consequential commitments that other agents may *attribute* to the agent, we avoid having to say that agents "believe" all logical consequences of their beliefs, including tautologies;

2. The correlate of *intentions* is taken to be (acknowledged) *practical* commitments; just as we can and do take on incompatible practical commitments without this licensing other agents to attribute arbitrary intentions to us so we can adopt incompatible *doxastic* commitments without thereby being committed to arbitrary propositions; however, we cannot be *entitled* to incompatible commitments.

MAS analysts have generally given only cursory attention to literature on the philosophy of mind and language, but this brief survey suggests that there are in fact sound philosophical arguments for conceptualising multi-agent interactions on the basis of operational notions of commitment rather than endowing agents with mental states.

1.2 From folk psychology to folk sociology?

In speech act theories, notions of *belief* and *intention* are generally taken as primitive and employed with their everyday senses; thus it could be argued that speech act theory rests on *folk psychology*. In a textbook presentation of multi-agent communication, Wooldridge points out a paradox in a speech-act based semantics for the inform locution in communication between autonomous software agents:

> If I am *completely* autonomous, and exercise *complete* control over my mental state, then *nothing* you say will have any effect on my mental state… if you are attempting to inform me of some state of affairs, then the best you can do is convince me that *you* believe this state of affairs.
>
> (Wooldridge, 2000a, pp. 137, emphasis in original)

This intuition is formalised in Wooldridge's LORA language by treating inform as an attempt to bring about a mutual belief in some group of hearers to the effect that the speaker intends them to mutually believe the asserted proposition (which, by the bye, does not seem to entail that the speaker believes ϕ). Formally (where i is an agent, g a group of agents, α an action, $\phi,\ \psi$ etc propositions):

> $\{\mathsf{Inform}\ i\ g\ \alpha\ \phi\} \mathrel{\hat{=}} \{\mathsf{Attempt}\ i\ \alpha\ \psi\ \chi\}$
>
> where
>
> $\psi \mathrel{\hat{=}} (\mathsf{M\text{-}Bel}\ g\ \phi)$
>
> and
>
> $\chi \mathrel{\hat{=}} (\mathsf{M\text{-}Bel}\ g\ (\mathsf{Int}\ i\ (\mathsf{M\text{-}Bel}\ g\ \phi)))$.
>
> In other words action α counts as an inform action between agent i (speaker) and group g (hearers) with respect to content ϕ if α is an attempt by i to cause the hearers to mutually believe ϕ, by at the very least causing g to mutually believe that i intends that g mutually believe ϕ.
>
> (Wooldridge, 2000a, pp. 137)

Thus there is no claim that an occurrence of $\{\mathsf{Inform}\ i\ g\ \alpha\ \phi\}$ *causes* any member of g to believe ϕ.

Now, it is not at all clear that this roundabout approach succeeds in resolving the "paradox of communication". If the hearer agent is "*completely* autonomous" then convincing it *that you believe* ϕ or even *that you intend it to believe* ϕ seem just as problematic as getting it to believe ϕ itself: both of these are still an attempt by the speaker to bring about a change in the hearer's mental state. One way out of this may be to circumscribe agents' autonomy: a member of an agent society should be expected to conform to certain minimal norms such as "if an agent i utters 'ϕ', assume that i intends you and other addressees to believe ϕ, and that the other addressees will make this assumption". Agents conforming to this norm would thereby be committed to automatically adjusting their mental state following utterances by other agents, in such a way as to secure the success of the weakest goal of Wooldridge's $\mathsf{Attempt}$ action above: "causing g to mutually believe that i intends that g mutually believe ϕ".

This (admittedly brief) argument strongly suggests that communication has an ineluctably *normative* dimension: agents need to share certain minimal conventions if communication is to get off the ground at all[1]. At this point, it may be instructive to come at things from the other direction and see whether normative terminology can be made *sufficient* as well as *necessary* for specifying the Inform action: that is, we will start by postulating various commitments associated with utterances, in the spirit of Brandom (1983, 1994, 2000) and see how far we can get without appealing to mentalistic notions of belief and intention.

The minimal norm we will assume as recognised by all rational agents is:

Norm 1 (Commitment) *to assert a proposition counts as taking on a commitment to subsequently defend the proposition*

(cf Walton and Krabbe, op. cit.).

This is an immediate consequence of defining a rational agent as one that not only has reasons for its actions but is capable capable of articulating reasons for those actions – including speech acts – rather than simply as a utility maximizer (as assumed in decision theory). Suppose agent i asserts ϕ in the presence of group g and "normal input-output conditions obtain" (Searle, 1965); the *normative effect* can be specified as:

(i) i publicly acknowledges commitment to ϕ, that is i undertakes to justify ϕ if challenged;

(ii) For all agents $j \in g$: j is both committed and entitled to attribute to i commitment to ϕ; that is, j

[1] For an early critique of the "instrumental" model of communication, see Habermas (1984a)

should be disposed to assent to the proposition *i is committed to ϕ* and to defend this commitment by uttering "i said that ϕ".

(iii) For all agents $\langle j, k\rangle \in g$: j is both committed and entitled to attribute to k commitment to i's commitment to ϕ; that is, j should be disposed to assent to the proposition *k is committed to i's commitment to ϕ* and to defend this commitment by uttering "k heard i say that ϕ".

(iv) and so on, for sequences of arbitrary length $\langle j \ldots n\rangle \in g$. E. g., "$k$ saw l hear i say that ϕ"

Clause (iv) implies an infinite nesting of commitments and may give the impression that agents are liable to indefinite paralysis as a consequence of trying to keep track of what they are committed to. However, this worry should dissolve if we keep in mind that commitment is interpreted here as a *social*, "deontic" status rather than a subjective psychological state. Agent i is committed to ϕ just in case some other agent j is entitled (according to social norms) to hold i to account and require the commitment to be redeemed.

The next step is a norm for *entitlement*:

Norm 2 (Entitlement) *If one succeeds in carrying out one's justificatory responsibility for [a claim] – if one succeeds in the dialogical game of giving and asking for reasons – other members of the linguistic community are bound to grant one entitlement to the claim.*

(Lance and Kremer, 1994, p. 372).

Now, the rules of the "dialogical game" will vary according to the type of dialogue. As Brandom observes:

> . . . generally, when a commitment is attributed to an interlocutor, entitlement to it is attributed as well, by default. (Brandom, 1994, p. 177).

That is to say, in the normal run of things we do not demand that all assertions be justified, but rather tend to take things on trust and assume that people have good enough reasons for what they say. Thus if an assertion goes unchallenged, not only can the assertor assume entitlement to commit to the asserted proposition, but hearers also become committed to the same proposition. In this case the hearers' commitment can be redeemed by *deferral*, e.g. by saying "i said that ϕ". An intermediate position would be where assertions are typically accepted without challenge if the assertor has a reputation for reliability or has privileged status within a community (this is further developed in section 2.2).

Perhaps the most radical feature of Brandom's system is the notion of *material inference*:

Norm 3 (Inference) *Inferential moves in discourse are governed by social practice in a community.*

For example: if John tells Mary he has been bitten by a Jack Russell terrier and Mary tells Bill "John was bitten by a dog", she is conforming to a socially-sanctioned regularity in her speech community such that when one calls something a Jack Russell terrier, one may also call it a dog. To put it another way: commitment to "x is a JR terrier" carries with it commitment to "x is a dog". Brandom argues that these "material", content-based inferences are primitives of linguistic competence and have explanatory priority vis-à-vis formal (logical) reasoning. Brandom's position is that formal rules for inference are merely *expressive* of the regularities observed by all norm-conformant agents in a speech community. In the context of MAS, such content-based inferences may be licensed by public ontologies as part of the semantic web; one may envisage emergent conventions such that agents converge on recognising the authority of particular ontology servers. The ontologies themselves can be seen as expressive of terminological norms within a particular specialist community (e.g., medical practitioners) but will take on a *normative* function when cited by web agents to fix the interpretations of predicates.

For anyone who is not clairvoyant, the only way to figure out what someone believes is by taking note of what they say, either spontaneously or in response to questions, and inferring any consequences or presuppositions of what they say. And the main reason for figuring out what someone believes is to be able to predict what they are likely to do in a given situation and what propositions they are likely to assent to or challenge. The working hypothesis for this paper, following Brandom, is that the "commitment/entitlement" framework can serve this purpose just as well as the "belief" framework, and as stated above, is actually more appropriate for modelling artificial agent interactions. One argument for this point of view is that it seems quite reasonable to attribute commitments to artificial agents even if they do not themselves have an articulated notion of "commitment"; that is, the commitment-entitlement framework can support reasoning about heterogenous agents, including BDI agents.

1.3 Desiderata

Having appeared to disparage the BDI approach as "folk psychology", it may well appear that we have instead taken refuge in "folk sociology", appealing to rather vague notions of what agents are "committed" to and what they may claim to be "entitled" to. It is therefore essential to define these terms sufficiently precisely that agents can unambiguously determine:

1. what commitments are in force at any time

2. what actions are needed to fulfil a commitment

3. when a commitment counts as *fulfilled*, *retracted* (by the "debtor"), *waived* (by the creditor) or *violated*

4. how commitments can be "policed" by the community, i.e. what *sanctions* are available in cases of unfulfilled commitments.

This paper will concentrate on the first question, which on its own poses some challenging problems. All four questions have already been tackled in various ways in the MAS literature e.g., Singh (1999); Fasli (2002); Yolum and Singh (2003); Parsons and McBurney (2003); Fornara and Colombetti (2003). (However, we should keep in mind that most work in MAS has focussed on practical commitments (commitments to action) rather than applying the concept to multi-agent communication.)

This question can be broken down into the following, among others (many of these issues were already noted by Hamblin (1970)):

- if I am committed to ϕ, which implications of ϕ can you hold me to be also committed to?

- if I retract commitment to ϕ, which other consequential commitments are also renounced? (A classic scenario in belief revision is that an agent may believe p and $p \rightarrow q$ and by inference, q; suppose q is subsequently retracted, consistency requires that either p or $p \rightarrow q$ be also retracted; but which?)

- if you grant my entitlement to commit to ϕ, which implications of ϕ can I expect you to grant me?

- if I retract an entitlement claim, which other entitlements are consequentially given up?

- does the same notion of "implication" apply in each case?

The remainder of this paper aims to elucidate the notions of commitment and entitlement to the point where we can begin to tackle these questions, and concludes by sketching a set of multi-agent update rules exploiting these notions. As a preliminary I offer a survey of different approaches to the concept of commitment itself.

2 Commitment and entitlement

As noted above Brandom makes the following claim

> The proposal is . . . not to *analyze* belief in terms of commitment but to discard that concept as insufficiently precise and replace it with clearer talk about different sorts of commitment.
>
> (Brandom, 1994, p. 196, emphasis in original)

The question arises, of course, whether the notion of commitment is any more precise than that of belief; in fact one finds various different uses and interpretations of this term in the literature.

2.1 Varieties of commitment

MAS commitments: Singh, Colombetti et al, Parsons & McBurney, Fasli

- commitments are specified as relations with the roles: *debtor* who takes on a commitment to a *creditor*, possibly in a *context* of third parties who do not play a direct part.
- each agent has a unitary, public commitment-store
- There are notional sanctions for violating commitments, though these are not spelled out
- commitments may be *conditional*, i.e. only come into force depending on stipulated events or states of affairs
- a variety of speech acts is supported: for example Singh (1999) gives definitions for inform, request, promise, permit, forbid, declare.
- commitments can be retracted via a cancel method - by the debtor only (Singh) or the creditor only (Fornara and Colombetti)
- there is no systematic notion of commitment entailments: this is (theoretically) problematic for retraction

- implementable via event calculus (Yolum and Singh, 2003), OOP (Fornara and Colombetti, 2003)

Brandomian commitments: Brandom, Lance and Kremer

- no explicit roles are specified: commitments are effectively made "to whom it may concern"...
- agents' commitment stores (deontic scoreboards) are perspectival, and distinguish other agents' acknowledged commitments and entitlement claims from commitments and entitlements which the scorekeeper attributes to them
- sanctions are assumed for nonfulfilment of a commitment (Brandom, 1994, p. 163) though apparently not for arbitrarily withdrawing commitments (Brandom, 2000, p. 93).
- there are elaborate discussions of committive/permissive entailment
- however, there is no formal account of *retraction* of commitments or entitlement claims
- only assertion (inform) is dealt with.
- not implemented, though partial formalisations have been proposed by Lance and Kremer.

Argumentation theory: Hamblin, Walton 93 etc, Walton & Krabbe 95

- commitments are classed as *light-side* (acknowledged), *dark-side* (unacknowledged) and *concessions* (commitments which an agent is not prepared to justify, but grants to an interlocutor "for the sake of argument"). (Walton and Krabbe, 1995)
- some attention to issues of entailment, retraction (esp.Hamblin)
- some discussion of *sanctions* (Walton adn Krabbe)
- protocols cover a variety of dialogue acts
- has influenced work on computational dialogue modelling

Some common factors are the treatment of commitment as a social status rather than a subjective psychological state[2], and the assumption that there should be some kind of sanctions for failure to redeem commitments. However, the efficacy of sanctions will of course depend on an agent's values and preferences, which could make it problematic to try and apply a common regime to humans and artificial agents. For instance, humans are affected to varying degrees to sanctions such as *ridicule* and *ostracism*, neither of which would be intrinsically effective against artificial agents.

It seems desirable to extend the Brandomian version to cover the full range of MAS locutions, incorporating committive and permissive entailments, and to consider carefully whether and what kinds of sanctions should be applicable when propositional commitments are withdrawn. Neither Walton's nor Brandom's work is presented in explicitly computational terms, or with sufficient precision and formality to be straightfowardly implementable. Formal treatments deriving from this work include Munindar Singh's social semantics (Singh, 1999; Yolum and Singh, 2003) and the family of relevant and non-relevant commitment logics presented in a series of papers by Mark Lance and Philip Kremer (Lance, 1995, 2001; Lance and Kremer, 1994, 1996).

A common assumption underlying Brandom's and Walton's work is that propositional commitment can be treated as a special case of action commitment. As noted above, Brandom introduces the normative dimension of *entitlement*: speakers not only acknowledge commitment to propositions but claim to be entitled to those commitments; to challenge a propositional claim is to seek to undercut or rebut a claim to entitlement. In any multi-agent interaction, each agent A_n maintains a set of commitments and entitlements for each agent A_i (deontic scoreboard) as shown in Figure 1. Commitments can be classified into *practical* (commitments to act, corresponding

- $C_{Ack}(A_i)$ Commitments A_i acknowledges
- $C_{Attr}(A_i)$ Commitments A_n attributes to A_i
- $E_{Cl}(A_i)$ Entitlements A_i claims
- $E_{Attr}(A_i)$ Entitlements A_n attributes to A_i

Figure 1: Entries in agent A_n's deontic scoreboard for agent A_i.

[2] Though Fasli (2002) appears to reserve the term *commitment* for an obligation which the agent has expressed an intention to carry out, or which is a recognised consequence of overtly adopting a social role; what we would call unacknowledged commitments appear in her work as obligations *tout court*.

to *intentions* in mentalistic accounts) and *doxastic* (commitments to justify an assertion, corresponding to *beliefs*).

Singh (1999) claims to be following Walton and also Habermas in his commitment-based semantics for the ACL primitives inform, request, promise, permit, forbid and declare. He classifies commitments as *objective*, *subjective* and *practical*, corresponding to Habermasian *validity claims* (see next section). In fact Singh's notion of commitment seems rather heterogenous; in his treatment of the inform(x,y,p) locution, the notion of propositional commitment as a special case of action commitment seems to have been lost:

Objective: C(x,y,p) - agent commits *that p is true*

Subjective: C(x,y,xBp) - agent commits *that it believes p*

Practical: C(x,G,inform(x,y,p) $\leadsto$ p) - agent commits *that it has reason to know p*

Thus from each perspective the inform locution results in a "commitment that" rather than a commitment *to do* something, which departs from the notion of commitment as "what an agent is obliged *to do*" (Singh, 1999, emphasis added). It is not obvious how such a commitment is supposed to be redeemed. This echoes Brandom's critique of Searlean speech act theory:

> ”One prominent theorist [Searle - RJK] defines the assertion of . . . p as "an undertaking to the effect that p". One does not have to subscribe to the pragmatist project . . . to find this disappointing. What sort of an undertaking is this? What, exactly, is the effect?
>
> (Brandom, 1994, p. 173)

In a similar vein, Fornara and Colombetti (2003) treat the effect of the inform locution as committing the assertor "to the truth of what is asserted". Singh states that

> Every commitment to a putative fact can be verified or falsified by challenging that putative fact.
>
> (Singh op. cit.)

However neither of these approaches appears to explicitly specify what *actions* an assertor is committed to in the event of a challenge.

It would be more perspicuous to build this into the definition of commitment so that e.g. inform(x,y,p) can be glossed as "agent x undertakes to agent y defend proposition p against dialogical challenges". So propositional commitments are perhaps best handled as *conditional* commitments (Fornara and Colombetti, 2003; Yolum and Singh, 2003): the agent is only required to redeem the commitment if appropriately challenged by another agent. (Nevertheless, for the remainder of this paper I will write "α is committed to ϕ" as a convenient shorthand for "α is committed to justifying ϕ if challenged by an interlocutor".)

2.2 Entitlements and validity claims

Brandom's notion of *entitlement* is glossed by Lance and Kremer as follows:

> If one succeeds in carrying out one's justificatory responsibility for [an assertion] A - if one wins in the dialogical game of giving and asking for reaons - other members of the linguistic community are bound to grant one entitlement to the claim.
>
> (Lance and Kremer, 1994, p. 372)

In fact as noted above, Brandom's formulation is not quite so rigorous: the "dialogical game" need not be played to the bitter end, but only to the point where acceptance of a claim is more reasonable than challenging it; and in everyday discourse this point tends to be reached quite soon.

I propose to add some structure to Brandom's notion of entitlement by incorporating a variant of Habermas's threefold "validity claims" (*Geltungsansprüche*). (Cf also Winograd and Flores (1986); Singh (1999); McBurney and Parsons (2003).) Under the formulation in Habermas (1984b) utterances raise three simultaneous claims, which the speaker undertakes to defend: they must be true (*wahr*), sincere or truthful (*wahrhaftig*) and "right" or appropriate to social norms (*richtig*). Heath (2003) argues against the notion that every speech act raises all three claims, and proposes that Habermas's account can only be made coherent on the assumption that the validity claims are associated with different types of discourse: theoretical, practical and expressive. Space does not permit elaboration of this issue, though I will note that Habermas (1998) introduces a distinction between "weak" and "strong" communicative rationality, whereby the former involves only the truth and sincerity claims. Most instances of multi-agent communication in the current state of the art would probably count as weakly rational in this sense.

For now I will adopt an approximation to Habermas' scheme whereby entitlement to doxastic commitmentscan be challenged or defended under one of the following headings[3]:

Type 1. Content of an utterance can be challenged by asserting an incompatible proposition, **or** by asserting a proposition which is incompatible with a precondition or a consequence of the proposition. The latter strategy assumes the interlocutor will endorse the relevant inference as well as the content of the challenge. Defending the content of a propositional commitment may involve appeal to observations or by inference to more "basic" commitments. Both of these may of course be open to further challenge.

Type 2. Reliability (truthfulness) is claimed for the speaker and for the source of any commitments which are inherited by testimony. Reliability can be challenged by e.g. instancing occasions when the speaker has (wittingly or not) uttered falsehoods, by questioning their qualifications or by raising doubts over "normal input-output conditions" in Searle's sense. For example, "you couldn't have seen that, it was too dark/you're near-sighted . . . " etc.

Type 3. Status: utterances may depend for their acceptability on the speaker's social role: for instance when making an insurance claim we may need to appeal to statements provided by agents who are not only recognised as reliable but have an appropriate institutional status: police reports, medical certificates etc.

In practice there can be some overlap between these categories: for instance "Trust me, I'm a doctor" can be glossed as either "My formal training and experience equip me to make reliable judgments" (Type 2) or "My professional status exempts me from scrutiny by layfolk" (Type 3).

Bearing in mind the requirement that agents cannot be entitled to incompatible commitments, validity claims must be strictly ranked. Direct observation or mathematical proofs (Type 1) will carry greater weight than supposedly reliable reports from third parties (Type 2), where these conflict. Likewise, a pronouncement by a forensic scientist as to a cause of death will normally be accorded the special status of "expert testimony" in a court of law (Type 3) but is not immune to empirical challenge.

3 Entailments and scorekeeping

The discussion so far has been rather informal. The task of formalising the above proposals can be divided (like Gaul) into three parts:

1. Committive inference: I will assume Lance's relevance system Lance (2001); this needs to be extended (in future work) with a treatment of *retraction*. Lance's system incorporates notions of agents being committed or opposed to a proposition as primitives in the model theory; these are independent stances which can co-exist in one and the same agent.

2. Permittive inference is yet to be formalised.

3. Communication is modelled as *inter-agent transfer of commitment*, formalised as a set of dynamic update rules (extending the system presented by Kibble (2005)).

Some preliminaries:

1. Each agent in a dialogue keeps a score of commitments and entitlements for all participants, including itself.

2. Agents play one of three (dynamically assigned) roles at any given point in a dialogue: **Speaker**, **Addressee**, **Hearer** (not directly addressed).

3. For an agent α to assert ϕ is to acknowledge commitment to ϕ; other agents may also attribute consequential commitments to α.

4. Each agent maintains a list of *Trusted* agents, which will by default be treated as *entitled* to assertions except where this leads to inconsistency (Type 2 entitlement). These may be agents which are *certified* as reliable or may have been *learned* to be reliable through repeated interactions. If agent α is on β's Trusted list, α may also nominate γ as Trusted in interactions with β.

5. Within an agent society there may be a category of *Privileged* agents whose assertions automatically carry entitlement (again, modulo inconsistency); this corresponds to Type 3 entitlement.

6. We distinguish two possible dialogue regimes:

   - *Default and challenge* (qui tacet consentit): if Speaker asserts ϕ, Addressee is held to endorse (be committed to) ϕ unless they challenge ϕ at the next appropriate opportunity; this assumption does *not* apply to Hearers.

[3] These headings cut across the three types of *grounding* (experiential, formal and social) proposed by Winograd and Flores (1986).

- *Sceptical*: an agent is only committed to claims which they endorse either explicitly or implicitly (e.g., by using as a premise for an inference at the next opportunity).

Each of these regimes requires a mechanism for recording Addressee's *potential* commitments following Speaker's assertion and until Addressee takes a turn as Speaker. To simplify the exposition we will assume that there can only be one such outstanding commitment at any time.

7. Speaker's claim to be entitled to commit to ϕ may be accepted, rejected or *deferred* pending further evidence or argumentation by Addressee and Hearers (cf Kibble (2001)). This requires a mechanism for recording *provisional* entitlements (not formally modelled in this paper).

3.1 Formalism

Scoreboard

(one per agent)

$\langle Sk,\ Tr,\ Pr,\ \{Sc_1 \ldots Sc_n\}\rangle$

Sk Identity of Scorekeeper

Sc$_i$ Scorecard for agent A_i

Tr List of *trusted* agents

Pr List of *privileged agents*

Both **Tr** and **Pr** are *partially ordered*; there are degrees of trust and of privileged status. This is specified as:

- $A_i <_{Pr} A_j$ - A_i, A_j are in **Pr** and A_i outranks A_j
- $A_i <_{Tr} A_j$ - A_i, A_j are in **Tr** and A_i outranks A_j
- $A_i <_{Ent} A_j$:
  - $A_i \in Pr$ and $A_j \notin Pr$ or $A_i <_{Pr} A_j$
  - $A_i \in Tr,\ A_j \notin Pr$ and $A_j \notin Tr$ or $A_i <_{Tr} A_j$

Scorecard

$$\langle R(A_i), C_{Ack}(A_i), \langle C_{Attr}(A_i),\ PC(A_i)\rangle,$$
$$E_{Cl}(A_i),\ \langle E_{Attr}(A_i),\ PE(A_i)\rangle\rangle$$

$R(A_i)$ Role of agent A_i: **Sp** | **Ad** | **He**

$C_{Ack}(A_i)$ Commitments acknowledged by A_i

$C_{Attr}(A_i)$ Commitments attributed to A_i

$PC(A_i)$ A_i's pending commitments or $\top$

$E_{Cl}(A_i)$ Entitlements claimed by A_i

$E_{Attr}(A_i)$ Entitlements attributed to A_i

$PE(A_i)$ A_i's provisional entitlements (stack), initially $\top$.

Let $\vdash$ stand for some notion of committive entailment, $\dashv$ for incompatibility, [.] for update, ϕ_i for "agent A_i asserts proposition ϕ", $C_{A_i}\phi$ for "agent A_i is committed to ϕ" and $A_i \bot \phi$ for "agent A_i is opposed to ϕ"

3.2 I. Updates to A_l's scorecard for Addressee A_j

1. $C_{Ack}(A_j)[\phi_i] = C_{Ack}(A_j)$
2. $\langle C_{Attr}(A_j), \top\rangle[\phi_i] =$
   $\langle C_{Attr}(A_j) \cup \{C_{A_i}\phi\},\ \phi\rangle$
3. $E_{Cl}(A_j)[\phi_i] = E_{Cl}(A_i)$
4. $E_{Attr}(A_j)[\phi_i] = E_{Attr}(A_j) \cup \{C_{A_i}\phi\}$

Clause I.2 has the effect that A_j is provisionally committed to ϕ and immediately committed and entitled to "A_i is committed to ϕ".

3.3 II. Updates to A_l's scorecard for Hearer A_k

1. $C_{Ack}(A_k)[\phi_i] = C_{Ack}(A_k)$
2. $\langle C_{Attr}(A_k), x\rangle[\phi_i] =$
   $\langle C_{Attr}(A_k) \cup \{C_{A_i}\phi\},\ x\rangle$
3. $E_{Cl}(A_k)[\phi_i] = E_{Cl}(A_k)$
4. $E_{Attr}(A_k)[\phi_i] = E_{Attr}(A_k) \cup \{C_{A_i}\phi\}$

The only effect on a Hearer who is not directly addressed is commitment and entitlement to "A_i is committed to ϕ".

3.4 III. Updates to A_l's scorecard for Speaker

1. $C_{Ack}(A_i)[\phi_i] = C_{Ack}(A_i) \cup \{\phi\}$
2. $\langle C_{Attr}(A_i), \chi\rangle[\phi_i]$
   $= \langle C_{Attr}(A_i) \cup \{\phi, \chi\} \cup \Sigma, \top\rangle$ **where**
   (a) $C_{Attr}(A_i) \cup \{\phi\} \not\dashv \chi$;
   (b) [*Sceptical*: $C_{Attr}(A_i) \cup \{\chi\} \vdash \phi$ ***and*** $C_{Attr}(A_i) \nvdash \phi$]

(c) $\forall \psi \in \Sigma,\ C_{Attr}(A_i) \cup \{\phi, \chi\} \vdash \psi$

else
= $\langle C_{Attr}(A_i) \cup \{\phi\} \cup \Sigma, \top \rangle$ **where**
$\forall \psi \in \Sigma,\ C_{Attr}(A_i) \cup \{\phi\} \vdash \psi$

3. $E_{Cl}(A_i)[\phi_i] = E_{Cl}(A_i) \cup \{\phi\}$

4. $E_{Attr}(A_i)[\phi_i] =$
$E_{Attr}(A_i) \cup \{\phi\}$ **if**
$\neg \exists \Sigma \subseteq (E_{Attr}(A_i) \cup C_{Ack}(A_l)) : \Sigma \dashv \phi$ **and**
$\neg \exists A_m(A_m \leq_{Ent} A_i \& A_m \bot \phi$
else
$= E_{Attr}(A_i)$

Clause III.2 covers attribution of consequential commitments: it does not say that *all* such commitments are added to the scoreboard but that any commitments added must be licensed by committive entailment, which is employed as a *filter* rather than a *generator* of commitments. This is because we want the scoreboard to remain finite.

This clause also says what happens to any provisional commitment χ added to the store in A_i's most recent turn as Addressee. If the regime is sceptical, χ becomes a persistent commitment only if A_i explicitly agrees with it, or utters some ϕ which assumes χ as a premise. If the regime is "default and challenge", χ becomes persistent if it is consistent with the current utterance ϕ.

Clause III.4 covers *entitlement*, which on this account is attributed by default unless the asserted proposition is inconsistent with the speaker's existing entitlements or the scorekeeper's commitments, or it is challenged by a higher-status or more trusted agent. Evidently the system needs to allow for retraction of entitlement attributions in case of *subsequent* challenges by higher-ranked agents.

4 Conclusions and future work

The following summarises the key claims that have been made in this paper as well as outstanding issues to be approached in future work.

- The folk-psychological notion of belief does not provide a firm foundation for computational semantics of dialogue acts, since we cannot reliably attribute beliefs to participants in these acts (the same goes for *intentions* and *desires*, but that's another story…). Furthermore, the notion that speech acts can *cause* changes in mental states appears incompatible with agent autonomy.

- Participants in dialogue are, however, expected to conform to norms and protocols which support the attribution of intersubjectively observable commitments to agents, and on the basis of which agents can claim entitlement to commitments.

- Brandom's "perspectival" model of commitment, entitlement, acknowledgement and attribution provides a more promising framework than the conventional notion of a single, public "commitment store" for each interlocutor. One benefit is that this allows for the possibility that agents use different inference regimes when reasoning about their own or other agents' commitments as opposed to entitlements, so that inconsistencies can be tolerated in the former case but not the latter. It is probably fair to say that paraconsistent logics are still viewed with suspicion in some quarters, though these suspicions may be somewhat allayed if the paraconsistency can be quarantined.

- Lance and Kremer's relevant and non-relevant commitment logics make it possible to formalise subtly different notions of commitment and to model agents' attributions of commitment to each other. From the point of view of computational implementation, solutions involving relevant entailment are somewhat unattractive and it may be worthwhile investigating whether the "desiderata" of the introductory section can be achieved by other means. In this paper L & K's work has been discussed in conceptual rather than formal terms and more rigorous analysis will be needed to determine whether their work can form the basis of a tractable implemented system.

- Formal properties of entitlement and permissive entailment have hardly been addressed in this paper; future work will develop the connection between Brandom's notion of entitlement and Habermasian validity claims, and will formalise a notion of defeasible permissive entailment.

- Brandom's notion of material inference as embedded in social practice has been somewhat under-emphasised in this paper; it remains unclear whether this notion can be carried over to societies of autonomous software agents. An alternative research programme would be to attempt to detach the framework of commitment, entitlement, acknowledgement and attribution from its Brandomian setting and redefine it in terms of more familiar AI approaches to knowledge representation and ontologies.

- The update rules presented above are somewhat rudimentary, as they apply only to positive atomic propositions and to "third-party" scorekeeping. Desirable extensions include:
  - Discuss implications for update rules when **Sk** = **Sp** or **Ad**
  - Extend update rules to handle Boolean operations and first-order formulas
  - Speech acts other than assertion:

    ϕ? - question: does hearer acknowledge commitment to ϕ; what commitments are undertaken in *asking* a question?

    ϕ! - imperative: speaker bestows *practical* commitment to the action described by ϕ on hearer

    $\phi\downarrow$ - retraction (downdate): speaker disavows commitment to ϕ.

Finally, I hope to have made a plausible case that the programme of formalising philosophical notions of propositional commitment and entitlement to inform agent design will lead both to a deeper understanding of these notions and to a productive framework for reasoning about how agents interact with each other and with their human clients.

Acknowledgements

I'm grateful to participants in the Sixth International Workshop on Computational Semantics, University of Tilburg, for useful feedback on an earlier presentation of these ideas.

References

J. L. Austin. *How to do things with words*. OUP, 1962.

Robert Brandom. Asserting. *Noûs*, 1983.

Robert Brandom. *Making it Explicit*. Harvard, 1994.

Robert Brandom. *Articulating Reasons: An Introduction to Inferentialism*. Harvard, 2000.

Philip Cohen and Hector Levesque. Rational action as the basis for communication. In *Intentions in Communication*. MIT, 1990.

Philip Cohen and Raymond Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 1979.

Maria Fasli. On commitments, roles and obligations. In *From Theory to Practice in Multi-Agent Systems, Second International Workshop of Central and Eastern Europe on Multi-Agent Systems, CEEMAS 2001*, 2002. Revised Papers. LNAI: 2296, Springer-Verlag, pp. 93-102.

Nicoletta Fornara and Marco Colombetti. Defining interaction protocols using a commitment-based agent communication language. In *AAMAS03*, 2003.

H P Grice. Meaning. *Philosophical Review*, 1957.

Jürgen Habermas. Intentionalistische Semantik (1975/6). In *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns*. Suhrkamp, 1984a.

Jürgen Habermas. Was heißt Universalpragmatik? (1976). In *Vorstudien und Ergänzungen zur Theorie des kommunikativen Handelns*. Suhrkamp, 1984b.

Jürgen Habermas. Some further clarifications of the concept of communicative rationality. In Maeve Cooke, editor, *On the Pragmatics of Communication*. 1998.

Charles Hamblin. *Fallacies*. Methuen, 1970.

Joseph Heath. *Communicative Action and Rational Choice*. MIT Press, 2003.

Rodger Kibble. Inducing rhetorical structure via nested update semantics. In *Proceedings of the Fourth International Workshop on Computational Semantics*, University of Tilburg, The Netherlands, 2001.

Rodger Kibble. Elements of a social semantics for argumentative dialogue. In *Proceedings of the Fourth Workshop on Computational Modelling of Natural Argumentation*, Valencia, Spain, 2004.

Rodger Kibble. Reasoning about commitments in dialogue. In *Proceedings of the Sixth International Workshop on Computational Semantics*, University of Tilburg, The Netherlands, 2005.

Mark Lance. Two concepts of entailment. *Journal of Philosophical Research*, 1995.

Mark Lance. The logical structure of linguistic commitment III: Brandomian scorekeeping and incompatibility. *Journal of Philosophical Logic*, 2001.

Mark Lance and Philip Kremer. The logical structure of linguistic commitment I: Four systems of non-relevant commitment entailment. *Journal of Philosophical Logic*, 1994.

Mark Lance and Philip Kremer. The logical structure of linguistic commitment II: Systems of relevant commitment entailment. *Journal of Philosophical Logic*, 1996.

David Lewis. *Convention*. Blackwell, 1969.

Peter McBurney and Simon Parsons. Engineering democracy in open agent systems. In *Engineering Societies in the Agents World (ESAW-2003)*, 2003.

Simon Parsons and Peter McBurney. Argumentation-based communication between agents. In Marc-Philippe Huget, editor, *Communication in Multiagent Systems*. Springer, 2003.

John Searle. What is a speech act? In M. Black, editor, *Philosophy in America*. Cornell University Press, 1965.

Munindar Singh. A social semantics for agent communication languages. In *Proc. IJCAI'99 Workshop on Agent Communication Languages*, pages 75–88, 1999.

Robert Sugden. The logic of team reasoning. *Philosophical Explorations*, pages 165–181, 2003.

Douglas Walton. Commitment, types of dialogue and fallacies. *Informal Logic*, 1993.

Douglas Walton and Eric Krabbe. *Commitment in dialogue*. SUNY Press, 1995.

Terry Winograd and Fernando Flores. *Understanding Computers and Cognition*. Addison-Wesley, 1986.

Michael Wooldridge. *Reasoning about rational agents*. MIT, 2000a.

Michael Wooldridge. Semantic issues in the verification of agent communication languages. *Journal of Autonomous Agents and Multi-Agent Systems. 3(1)*, pages 9–31, 2000b.

Pinar Yolum and Munindar P. Singh. Reasoning about commitments in the event calculus: An approach for specifying and executing protocols. *Annals of Mathematics and Artificial Intelligence (AMAI), Special Issue on Computational Logic in Multi-Agent Systems*, 2003.

Formalizing Coalition Structures via Obligations

Luigi Sauro[*]
[*]Dipartimento di Informatica - Università di Torino
Corso Svizzera 185 - Torino
sauro@unito.it

Abstract

In this paper we present a definition of admissibility for coalitions based on a balance between the advantages and the obligations an agent is subject in signing a contract. We also define a notion of profitability and we show that admissible coalitions satisfies this notion. A relevant aspect of the definition of admissibility is that it is not based on an esplicit utility function associated to the goals of an agent. It simply resort on the fact that an agent prefers to receive a set of goals G_1 to a set of goals G_2 if G_1 contains G_2, on the contrary he prefers to provide a set of goals G_3 to a set of goals G_4 if G_3 is contained in G_4.

1 Introduction

In recent years an increasing interest has been focused on the formalization of normative systems. In particular, a novel problem regards the correct description of the dynamics under which different individuals may change their right-obligation status.

As described in Boella & van der Torre (2004), contracts constitute legal institutions that can modify the set of obligations of the participating agents.

The topic of this work is not to provide a general description of what a contract is, but to characterize some aspects of the social function of contracts and to frame them in the context of multiagent systems. In particular a contract is viewed as a set of regulative rules that provide to the participating agents the possibility to exchange goods or to mutually satisfy goals that otherwise could not be satisfied. In this perspective contracts allow the formation of dynamic alliances, Sandholm & Lesser (2002), or, as we denote in this work, coalitions. The main difference with respect to the usual regulative norms is that contracts enable collusive behaviors, so it is required that all the participating agents agree the contract in order to make it operative.

In this context we consider the decision-making processes any agent should carry out in order to sign profitable contracts. This issue has been largely discussed in several works based on well known equilibrium criteria borrowed from game theory Kraus & Schechter (2003); Shehory & Kraus (1998); Sandholm & Lesser (1997, 2002). Referring to Sandholm & Lesser (1997) this process is composed in two phases. In the first phase a coalition structure, describing the set of all admissible coalitions, is considered. The second phase concerns the selection of a subset of the coalition structure that satisfies a solution criterion, as for example the Nash equilibrium or the core determination, or that maximize the social welfare Sandholm & Lesser (1997).

Some works, such as Sandholm & Lesser (1997) and Klusch & Shehory (1996), specify very simple properties for a coalition structure, for example the set of partitions of the agents. Therefore a coalition structure does not contain any particular information about the profitability of coalitions. On the contrary we use the description of a contract to find out some qualitative conditions that allow to cut off unprofitable coalitions from a coalition structure.

The profitability of a coalition is defined by means of two conditions. The first is that an agent, involved in a contract, does not agree to satisfy a goal if it is not useful for the satisfaction of one of his own goals. The second is related to the fact that different coalitions can exist independently, and the activity of one cannot influence the formation of the others. We argue that aggregating independent coalitions can not be considered a coalition and that, in this case, it is more profitable if two distinct contracts occur. We refer to this issue as the compositional problem and we address it by identifying when given two admissible coalitions, their union is an admissible coalition.

In this work we consider a simple description of a contract as a relation that associates sets of agents to goals they are obliged to satisfy in the case the contract is operative; a contracting party is a set of agents and an obligation is intended as a goal the party has the duty to accomplish.

This formalization is subject to some assumptions. First, we do not consider the possibility that an agent can decommit a signed contract as in Sandholm & Lesser (2002). Second, we assume that the agents are completely confident in the efficiency of the normative system in detecting and punishing violations, moreover an agent always prefers to accomplish an obligation with respect to the relative sanction.

In Section 2 we provide a definition of admissibility for coalitions. In Section 3 we show how admissible conditions are profitable. Finally in Section 4 conclusions and future works are delineated.

2 Obligation-based admissible coalitions

In our framework a coalition is a set of agents that can exchange goals with each other. The exchanges are ratified by a contract, therefore a possible coalition is described by means of a contract.

Not all the coalitions are profitable for a participating agent, for example a coalition in which an agent provides something without receiving anything is not profitable. Moreover, even if two coalitions are individually profitable, their union could not be profitable (composition problem).

As an example of the composition problem, let assume the following scenario: there are four collectors a_1, a_2, a_3 and a_4; a_1 and a_2 are stamp collectors, whereas a_3 and a_4 car collectors. Now, a_2 has a rare stamp that a_1 desires and vice versa a_1 has a stamp desired by a_2. The same thing occurs for a_3 and a_4 with cars.

Assume now that all of them are invited to sign a single contract that binds them to exchange their goods. The contract is operative only if all the collectors decide to sign it. If, for example, the collector a_3 does not consider fair the exchange with a_4, then he does not sign the contract and hence the entire deal is upset. Therefore, for a_1 and a_2 it has not been profitable to consider their deal together with a_3 and a_4, being so dependent on their agreement. The entire contract can be separated in two independent subcontracts, the first relative only to the exchange between a_1 and a_2, and the second relative only to the exchange between a_3 and a_4. In this case a_1 and a_2 are no more dependent on a_3 and hence they can exchange their stamps.

In this section we provide a definition of admissibility for coalitions based on a balance between the goals an agent could obtain and the ones he is obligated to satisfy in the case the corresponding contracts become operative. We show in the next section how this definition addresses the problem to describe profitable coalitions.

First, we consider a set of agents $Ag \equiv \{a_1, \ldots, a_n\}$, a set of goals Gl and a function $gl : Ag \rightarrow 2^{Gl}$ that associates, to any agent a_i, the set of his goals. Now we formalize the specification of a contract. As said in the previous section, we describe a contract as a relation e that associates to a set of agents a set of goals, the idea is that if the goal g is assigned to a set of agents A, $(A, g) \in e$, then, when the contract become operative, all the agents in A are obligated to provide, as a group, g.

So we have the following definitions:

Definition 1 (Contracts) *A contract is a relation* $e \quad 2^{Ag} \quad Gl$*. Moreover, provided the denotations:*

1. $e(A) = \{g \in Gl \mid (A, g) \in e\}$
2. $Dom(e) = \{a \in Ag \mid \exists A \quad Ag \; a \in A \wedge e(A) \neq \emptyset\}$

We call any $(A, g) \in e$ *an assignment of* A *and* $Dom(e)$ *the domain of* e.

Given a contract e, we also define a two functions, $adv_e : Ag \rightarrow 2^{Gl}$ and $obl_e : Ag \rightarrow 2^{Gl}$. The first returns, for a given agent, all the goals he benefit from e. The second returns all the goals he is involved in an assignment. More formally the function adv_e is defined as:

Definition 2 (adv function) *Given a contract* e, $adv_e : Ag \rightarrow 2^{Gl}$ *is such that for all* $a \in Ag$

$$adv_e(a) \equiv \{g \in gl(a) \mid \exists A \quad Ag \; g \in e(A)\}$$

Instead, the function obl_e is defined as:

Definition 3 (obl function) *Given a contract* e, $obl_e : Ag \rightarrow 2^{Gl}$ *is such that for all* $a \in Ag$

$$obl_e(a) = \{g \in Gl \mid \exists A \quad Ag \; a \in A \wedge g \in e(A)\}$$

The functions adv_e and obl_e describe, for any agent, the balance between advantages and obligations with respect to the contract e.

We define an admissibility condition for coalitions by means of contracts, and hence without referring to any explicit description of the utilities and costs associated to the goals. Therefore, we define a preference relation such that it is not possible to directly compare two goals, because we do not know their utilities. Nevertheless an agent should aim to maximize the set of goals he recivies and to minimize the set of goals he is obliged to provide, so the idea is that, given two contracts e_1 and e_2, an agent prefers e_1 to e_2 if the set of goals he receives by means of e_1 contains the one he recivies by means of e_2 and the set of goals he is obliged to provide in e_1 is contained in the one he is obligated to provide in e_2.

Definition 4 (Preference relation) *Let* e_1 *and* e_2 *two contracts. We say that the agent* a_i *prefers* e_1 *to* e_2*,* $e_2 \quad {}_i\ e_1$*, iff* $adv_{e_2}(a_i) \qquad adv_{e_1}(a_i)$ *and* $obl_{e_1}(a_i) \quad obl_{e_2}(a_i)$

As usual we say that a_i strictly prefers e_1 to e_2, $e_2 <_i e_1$, if $e_2 \quad {}_i\ e_1$ and $e_1 \ /_i\ e_2$.

In particular, considered a contract e and a subset $e' \quad e$ we have that for all a_i

$$adv_{e'}(a_i) \quad adv_e(a_i) \wedge obl_{e'}(a_i) \quad obl_e(a_i) \qquad (1)$$

Therefore the following lemma holds:

Lemma 1 *Given* $e' \quad e$*,* $e \quad {}_i\ e'$ *if and only if* $adv_{e'}(a_i) = adv_e(a_i)$*.*

Provided a preference relation it is possible to define a dominance criterion. This criterion is defined by means of a relation *Dominates* between two contracts e and e' such that $e' \quad e$. The idea is that e is dominated by e' if all the agents in the domain of e' are not interested in the goals negotiated in $e \setminus e'$ or it exists an agent a_j such that he gives and provides something from both e' and $e \setminus e'$, but all the other agents in e' are neither interested in $e \setminus e'$ nor helpful for the agents involved in it and the same thing occurs for the agents in $e \setminus e'$. So, even if a_j takes a part in both e' and $e \setminus e'$, his decision to join $e \setminus e'$ cannot influence the decision of some a_i in e' to join e' and the same thing occurs with respect to the agents involved in $e \setminus e'$. So he can decide about e' and $e \setminus e'$ separately.

We first provide the formal definition of the dominance relation and after we have all the ingredients to define admissible coalitions.

Definition 5 (Dominance) *Given a contract* e*, we say that* $e' \quad e$ *dominates* e*, Dominates*(e', e)*, iff:*

$$\forall a_i \in Dom(e') \ \ e \quad {}_i\ e'$$

or it exists $a_j \in Dom(e')$ *such that:*

1. $e' \not<_j e$
2. $\forall a_i (\neq a) \in Dom(e') \ \ e =_i e'$
3. $\forall a_k \notin Dom(e') \ \ e =_k e \setminus e'$

Definition 6 (Admissible coalitions) *We say that a contract* e *is an admissible coalition iff*

$$\forall e' \subset e \ \ \neg Dominates(e', e)$$

Now we emphasize some properties of the previous definitions. First, the following lemma holds:

Lemma 2 *If in a contract* e *there are two distinct sets of agents* A_1*and* A_2 *binded in the assignment of the same goal* g*, then* e *is not an admissible coalition.*

proof:

Let e' the contract obtained from e removing (A_2, g). The set of goals assigned in e' is the same of those assigned in e. Therefore, by virtue of the Lemma 1, for all the a_i in $Dom(e')$, $e \quad {}_i\ e'$ and hence, by virtue of the first condition in the Definition 5, e' dominates e.$\diamond$

Assume now that e' dominates e, we have from Definition 5 two possible cases. In the first case for all the agents a_i in $Dom(e')$, $e \quad {}_i\ e'$. In this case, by virtue of Lemma 1, we have that $adv_{e \setminus e'}(a_i) \quad adv_{e'}(a_i)$. So if they form the coalition prescribed by e', then they are no more interested in the goals provided in $e \setminus e'$.

If the second condition of the Definition 5 holds, then there exists an agent $a \in Dom(e')$ such that for all $a_i(\neq a) \in Dom(e')$ and for all $a_k \in Dom(e \setminus e')$

$$adv_{e \setminus e'}(a) = \emptyset \quad \vee \quad obl_{e \setminus e'}(a) \neq \emptyset \qquad (2)$$

$$adv_{e \setminus e'}(a_i) = \emptyset \quad \wedge \quad obl_{e \setminus e'}(a_i) = \emptyset \qquad (3)$$

$$adv_{e'}(a_k) = \emptyset \quad \wedge \quad obl_{e'}(a_k) = \emptyset \qquad (4)$$

So the agents a_i in $Dom(e')$ are not interested in the goals provided in $e \setminus e'$ and do not participate in the achievement of any of the goals in it. Similarly, also the agents a_k in $Dom(e \setminus e')$ are not interested in the goals achieved in e' and do not participate in the achievement of any of the goals in it. Finally either a is not interested in $e \setminus e'$, and then also the previous case holds, or he have to provide something in $e \setminus e'$.

3 Profitability of coalitions

In this section we show how Definition 6 satisfies the profitability of coalitions.

The profitability of a coalition is based on two conditions:

> **Do ut des property:** Agents do not accept to be obliged to satisfy a goal if it is not useful in obtaining the satifaction of some of their own goals.

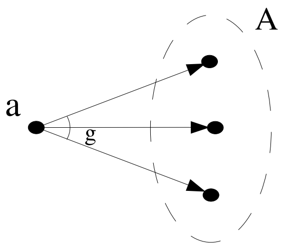


Figure 1: *And-arrow*

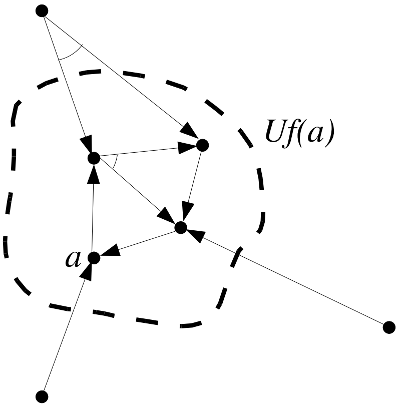


Figure 2: *A coalition that does not satisfy the do ut des property*

> **Composition problem:** If e is a profitable coalition, then it does not exist a profitable coalition $e'(\neq \emptyset) \subset e$ such that the formations of e' and $e \setminus e'$ are independent processes.

First of all we need to describe the previous conditions in terms of the goals of an agent and his obligations in a contract e.

To use figures we represent e as follows: if $(A, g) \in e$, then for all agents a that desire g we draw an and-arrow from a to A tagged with g (see Figure 1) and we say that the and-arrow starts from a and, for any agent in A, it reaches him. Thus, given a contract e and its graph representation, $adv_e(a)$ is the set of tags for which an and-arrow exists starting from a, similarly $obl_e(a)$ is the set of tags such that an and-arrow exists that reaches a. In the following, for simplicity, when we omit tags in the figures, then we assume that they are all different.

To formalize the do ut des property two preliminary definitions are required. First we say that, given an agent a and a contract e, the contract e' $\quad e$ is the directly useful contract of a, $Duf(a)$, if and only if e' is composed by all the pairs (A, g) such that g is a goal of a.

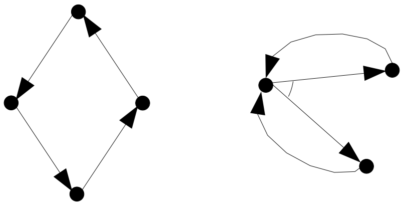


Figure 3: *A coalition that satisfies the do ut des property, but it does not satisfy the composition problem.*

Secondly, we define the contract e' $\quad e$ to be the useful contract for an agent a, $Uf(a)$, if and only if e' is the minimal set that contains the directly useful contract of a and the union of the direct useful contract of all the agents in the domain of $Duf(a)$ and so on (see Figure 2). $Uf(a)$ is the subset of e useful for a, i.e. the set of assignments he can use to realize a chain of exchanges in order to obtain the satisfaction of his own goals. The do ut des property establishes that in a profitable coalition all the assignments in which a is involved have to be useful for him; for example Figure 3 shows a coalition that satisfies the do ut de property.

Definition 7 (Directly useful contracts) *Given a contract e, the directly useful contract of an agent a, $Duf(a)$, is the subset of e:*

$$Duf(a) \equiv \{(A, g) \in e \mid g \in gl(a)\}$$

Definition 8 (Useful contract) *Given a contract e, the useful contract of an agent a, $Uf(a)$, is the minimal set that satisfies the following conditions:*

1. *$Duf(a) \quad Uf(a))$*
2. *if $a' \in Dom(Uf(a))$, then $Duf(a') \quad Uf(a)$*

Now we prove that an admissible coalition satisfies the do ut des property:

Theorem 1 *If a contract e is an admissible coalition, then, for any agent a and for any $(A, g) \in e$ such that $a \in A$, $(A, g) \in Uf(ag)$.*

proof:

Assume that it exists an agent a such that the consequent is false, so it exists a set of agents A, that contains a, such that $(A, g) \in e$, but $(A, g) \notin Uf(a)$. Clearly this entails that $Uf(a) \subset e$. We show also that $Uf(a)$ dominates e against the hypothesis that e is an admissible coalition (see Definition 6). In fact it follows from the Definition 7 that, for any agent a_i, the set of advantages relative to e, $adv_e(a_i)$ is equal to the set of advantages relative to his directly useful contract, $adv_{Duf(a_i)}(a_i)$, and hence, considering the Lemma 1, $e \quad _i\ Duf(a_i)$.

Since in Definition 8, by construction, for all $a_i \in Dom(Uf(a))$, $Duf(a_i) \quad Uf(a)$, we have that

$$adv_e(a_i) = adv_{Duf(a_i)}(a_i) \quad adv_{Uf(a)}(a_i) \quad adv_e(a_i)$$

and hence, using again the Lemma 1, for all $a_i \in Dom(Uf(a))$, $e \quad _i\ Uf(a)$. Therefore, the first condition of the Definition 5 is satisfied against the hypothesis that e is an admissible coalition.$\diamond$

Proved that an admissible coalition satisfies the do ut des property, the following lemma holds:

Lemma 3 *If e is an admissible coalition, then for all a in $Dom(e)$, $adv_e(a) \neq \emptyset$.*

Now we focus on the composition problem. The composition problem says that, given a profitable coalition e, it does not exist a profitable coalition $e_1 \subset e$, such that the formation processes of e_1 and $e \setminus e_1$ are mutually independent. The formation processes of e_1 and $e \setminus e_1$ are mutually independent if there does not exist two distinct agents, one in $Dom(e_1)$ and the other one in $Dom(e \setminus e_1)$, that are interested in both e_1 and $e \setminus e_1$. Indeed, if the set of agents in $Dom(e_1)$ interested $e \setminus e_1$ is empty, then they can stipulate e_1 separately from the agent in $Dom(e \setminus e_1)$, avoiding the problem of being dependent, as in e, on their agreement (see Figure 3). The same holds if all the agents in $Dom(e \setminus e_1)$ are not interested in e_1.

Moreover, even if there exists only one agent a in $Dom(e_1) \cap Dom(e \setminus e_1)$ which is interested in both of them, as in Figure 4, no other agent in e_1 is interested on his decisions about $e \setminus e_1$ and vice versa. So also a can consider the two coalitions as independent. On the contrary if there are two distinct agents, respectively in $Dom(e_1)$ and $Dom(e \setminus e_1)$, interested in both the contracts, then they can negotiate about their decisions to sign both or none of them.

Given a contract e we denoted with $Int(e)$, the set of agents interested to e, that is:

$$Int(e) = \{a \in Ag \mid adv_e(a) \neq \emptyset\}$$

The following theorem shows that the Definition 6 satisfies the composition problem.

Theorem 2 *If a contract e describes an admissible coalition, then for all admissible coalitions $e_1 \subset e$, denoted with $\mathcal{A} = Dom(e_1) \cap Int(e \setminus e_1)$ and with $\mathcal{B} = Dom(e \setminus e_1) \cap Int(e_1)$, the following conditions holds:*

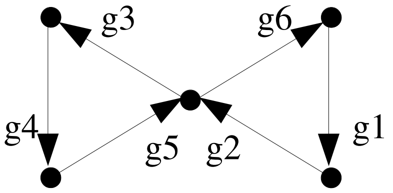


Figure 4: *A coalition that satisfies the do ut des property, but it does not satisfy the composition problem*

1. $\mathcal{A} \neq \emptyset$
2. $\mathcal{B} \neq \emptyset$
3. $|\mathcal{A} \cup \mathcal{B}| > 1$

proof:

Assume that there exists an admissible coalition $e_1(\subset e)$ such that one of the previous items does not hold. We prove that e_1 or $e \setminus e_1$ dominates e, against the hypothesis that e is admissible.

If $Dom(e_1) \cap Int(e \setminus e_1) = \emptyset$, then for all $a_i \in Dom(e_1)$, $adv_{e \setminus e_1}(a_i) = \emptyset$ and hence $e \quad _i\ e_1$. But this means that e_1 dominates e.

An analogous proof can be applied in the case $Dom(e \setminus e_1) \cap Int(e_1) = \emptyset$ showing that $e \setminus e_1$ dominates e.

In the last case $|\mathcal{A} \cup \mathcal{B}| \quad 1$. Now if $|\mathcal{A} \cup \mathcal{B}| = 0$, then both $\mathcal{A}$ and $\mathcal{B}$ are empty and we fall in the previous cases. So assume that $|\mathcal{A} \cup \mathcal{B}| = 1$ and, without loss of generality, that $\mathcal{A} \neq \emptyset$ and $\mathcal{B} \neq \emptyset$.

This means that $\mathcal{A} = \mathcal{B} = \{a_l\}$, so for all $a_i(\neq a_l) \in Dom(e_1)$, $a_i \notin Int(e \setminus e_1)$, and hence $e \quad _i\ e_1$. Analogously, for all $a_j(\neq a_l) \in Dom(e \setminus e_1)$, $a_j \notin Int(e_1)$ and hence $e \quad _j\ e \setminus e_1$.

Finally, we have that $a_l \in Dom(e \setminus e_1)$, and hence $obl_{e \setminus e_1}(a_l) \neq \emptyset$. Moreover, since by hypothesis e is admissible, by virtue of the Lemma 2, $obl_{e \setminus e_1}(a_l) \cap obl_{e_1}(a_l) = \emptyset$. But this means that $obl_{e_1}(a_l) \subset obl_e(a_l)$ and hence $e_1 \not\prec_l e$.

So we have proved that there exists an agent $a_l \in Dom(e_1)$ such that (1) $e_1 \not\prec_l e$, (2) for all $a_i(\neq a_l) \in Dom(e_1)$, $e \quad _i\ e_1$, (3) for all $a_j(\neq a_l) \in Dom(e \setminus e_1)$, $e \quad _j\ e \setminus e_1$. But this means that the second condition of the Definition 5 holds and hence e_1 dominates e.$\diamond$

4 Conclusion and Future works

In this work we provide a definition of admissibility for coalitions that is based on the balance between the goals an agent can obtain in signing a contract with respect to the ones he is obliged to satisfy. This admissibility condition differentiates with respect to the traditional criteria based on pay-offs in the fact that the goals are not mapped in a monetary counterpart by means of an utility function. Therefore, they are not directly comparable. From a mathematical point of view this entails that the preference relation of the agents is a partial order instead of a total order.

Moreover it has been also formalized a criterion of profitability for a coalition based on two issues: the do ut des property and the composition problem. By means of some examples it is shown the relevance of these two issues and it is proved that the admissible coalitions are profitable.

The addressed problem has been analyzed trying to simplify as much as possible the contextual framework. In this way it has been possible to individuate some relevant aspects that play a role in its solution abstracting from others that, even if could influence the admissibility of a coalition, not always occur or can have different weights. Example of these secondary aspects are, for example, the possibility to decommit a contract, or the trust on the monitoring normative system as in Boella & Lesmo (2002).

We have assumed that, in a contract, if a goal g is associated to a set of agents A then A has the power to achieve g, and also we have assumed that all the goals are compatible. In any case we have not focused on this aspect, so a possible development of this work consists in the relate our definitions of contracts and admissible coalitions to a notion of power and dependence as studied in a sociological context, Castelfranchi (2003), or in a game theoretical framework, Brainov & Sandholm (1999).

References

Boella, G., and Lesmo, L. 2002. A game theoretic approach to norms and agents. *Cognitive Science Quarterly* 492–512.

Boella, G., and van der Torre, L. 2004. Contracts as legal institutions in organizations o fautonomous agents. In *Procs. of AAMAS'04*.

Brainov, S., and Sandholm, T. 1999. Power, Dependence and Stability in Multiagent Plans. In *Procs. of AAAI'99*, 11–16.

Castelfranchi, C. 2003. The Micro-Macro Constitution of Power. *ProtoSociology* 18-19.

Klusch, M., and Shehory, O. 1996. A polynomial kernel-oriented coalition algorithm for rational information agents. In *Procs. of ICMAS'96*.

Kraus, S., and Schechter, O. 2003. Strategic Negotiation for Sharing a Resource between Two Agents. *Computational Intelligence* 19:9–41.

Sandholm, T., and Lesser, V. 1997. Coalitions among computationally bounded agents. *Artificial Intelligence* 94(1):99–137.

Sandholm, T., and Lesser, V. 2002. Leveled-commitment contracting. *AI Magazine* 23(3):89–100.

Shehory, O., and Kraus, S. 1998. Methods for Task Allocation via Agent Coalition Formation. *Artificial Intelligence* 101(1-2):165–200.

Impact Of Multiple Normative Systems
On Organization Performance Of International Joint Ventures

Tamaki Horii
Civil and Environmental Engineering
Department
Stanford University
Stanford, CA 94305, U.S.A
Tamaki_Horii@stnaford.edu

Yan Jin
Aerospace and Mechanical Engineering
Department
University of Southern California
Los Angeles, CA 90089-1453
yjin@usc.edu

Raymond E. Levitt
Civil and Environmental Engineering
Department
Stanford University
Stanford, CA 94305, U.S.A
ray.levitt@stanford.edu

Abstract*

Research on international joint ventures (IJV) reveals difficulties in managing cross-cultural teams. Our research aims to understand how cultural differences between Japanese and American firms in IJV projects effect team performance through computational experimentation. We focus on and characterize culturally-driven normative systems as a central role of cultural differences, composed of two dimensions: *cultural values* and *cultural practices*. *Cultural values* refer to workers' preferences in making task execution and coordination decisions. These preferences drive specific micro-level behavior patterns for individual workers. *Cultural practices* refer to workers' perceptions and expectations to include norms adopting each culture's typical organization style, such as centralization of authority, formalization of communication, and depth of organizational hierarchy. Our ethnographic observations have documented distinctive micro-level *behavior patterns* and *organization styles* for Japanese and American teams. We use a computational experimental design that sets *task complexity* at four levels and *team experience* independently at three levels, yielding twelve organizational contexts. We then simulate the four possible combinations of US *vs.* Japanese team *individual behavior* and *organization style* in each context to predict *work volume, cost, schedule,* and *project quality* outcomes. Simulation results predict that: 1) both Japanese and American teams show better performance across all contexts when each works with its typical organization style, suggesting positive correlation between two normative components (*cultural values* and *cultural practices*) on team performance; 2) the Japanese organization style performs better in the case of high task complexity, while the American organization style performs better in the cases of low and medium task complexities, implying that the impact of normative systems is contingent upon task complexity; and 3) the Japanese organization style tends to have significantly lower project quality (system integration) risks than the American organization style. In addition, *cultural practices* (typical organization styles) have a larger impact on project performance than *cultural values* (culturally driven behavior patterns). Our simulation results are qualitatively consistent with both organizational and cultural contingency theory, and with limited observations of US-Japanese IJV project teams.

1. Introduction

In an era of globalization, as economic borders between countries come down, cultural barriers will most likely appeare and present new challenges and opportunities in business (House et al, 2004). Projects in the construction industry face unique challenges in coordinating among sponsors, financiers, developers, designers and contractors from different countries. The project participants work for companies with varying corporate cultures and management styles. Their companies' headquarters are located in the different countries so that project teams need to overcome a variety of languages, business customs, and cultures. In addition, project products are mostly one-of-a-kind and largely handcrafted as the inherent nature of construction industry. Therefore, in order to building facilities, project teams generally need to cope with various pressures such as short project duration, strict budget, local institutions, and physical environments.

Research on international joint-venture (IJV) projects reveals significant difficulties in managing cross-cultural teams. According to one study, two out of every five IJV project teams struggle through their projects and show poor performance (Beamish and Delios, 1997). One key problem is the increased internal complexity caused by pre-existing differences among IJV team members in cultural values, beliefs, norms, and work practices. In particular, normative

* This research is being conducted under the auspices of the Collaboratory for Research on Global Projects (CRGP) <http://crgp.stanford.edu>. This material is based upon work supported by the National Science Foundation under Grant No. 9980109, the Clarkson H. Oglesby Memorial Fellowship Fund, and industrial affiliates of CRGP. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

systems including both values and norms (Scott, 2001) can define appropriate behaviors and legitimate processes, playing a central role in cultural differences between subgroups composing of an IJV project team. In other words, IJV projects can be viewed as a place where different normative systems bump into together, giving us a great opportunity to observe cross-cultural effects on team performance.

In order to lead a project successfully, project managers need not only to comprehend the differences in normative systems among their partners, but also to understand the influence of these differences on team performance. This research attempts to characterize differing normative systems that emerge in IJV teams, and to model and analyze effects of normative systems on team performance through computational "virtual" experimentation.

In this research, we focus on two cultures—Japanese and American—as an example of the minimum dyadic unit of cultural and institutional interaction in global construction projects. We expect that we can generalize our findings which will be applicable to other cultures. Many researchers have characterized distinguishing differences between Japanese and American cultures in business situations (e.g., Nakane, 1970; Ouchi, 1981; Aoki, 1992). Their findings help in understanding the internal consistency of Japanese *vs.* American social and organizational principles and their differences from one another. However, very few cross-cultural studies have focused on the construction industry, even though the international construction market alone is worth $106.5 billion (June 2001 issues of Engineering News Record (ENR) magazine$^{)}$.

We begin by defining culture. Generally, culture can be defined as a set of shared experiences, understandings, and meanings among members of a group, an organization, a community, or a nation (e.g., Redfield, 1948; Davis, 1984; Schein, 1989; Hofstede, 1991). Through sharing common successes and struggles, groups create their own unique cultures, leading to the development of unique sets of values—i.e., broad tendencies to prefer certain states of affairs over others—and practices (norms)—i.e., conceptions of appropriate business practices to include legitimate means and processes-. Both *cultural values* and *cultural practices* have been elaborated and fostered as culturally-driven normative systems of a social or a group for years, playing a central role of cultural differences in IJV projects. Therefore, this research views cultural differences from two dimensions: *value differences* and *practice differences* (Hofstede, 1991; House et al, 2004). Hofstede (1991) originally describes national culture in terms of both values and practices. Although our focus of this research is on project organizations rather than national culture, the dimensions of value and practices provide a good starting point for us to study culture and cultural differences in project teams. Our work extends Hofstede's definitions to cover project organizations.

Computer simulation is growing in popularity as a research method for organizational researchers (Dooley, 2002). Multi-agent based models, such as the Virtual Design Team (VDT) (Levitt et al, 1994; Jin and Levitt 1996), can provide a laboratory to address "what-if" questions about project team performance and organization design (Burton, 2003). The VDT model was not originally intended to capture cultural factors, but its rich characterization of both organizational and actor behaviors provide some capability to model cultural phenomena. The long term goal of this research is to extend the representation and reasoning of the extant VDT model to capture the impact of cultural differences in global construction projects. As the first step toward this goal, the current research explores the extent to which the VDT model can be used to model cultural influences on project team performance.

In this research, we take the following steps to analyze how culture impacts on project team performance. First, we characterize the typical normative systems of Japanese teams and American teams in terms of their value differences and practice differences, based on literature and our observations. Second, we encode selected cultural factors into the micro-level behavior and organizational parameters of the VDT model. Third, we analyze the effects of value and practice changes on team performance through "Intellective Simulation" using idealized organizations (Burton & Obel, 1995). Finally, the simulated results are qualitatively compared with "Cultural Contingency" propositions for the "preferred coordination mechanism[1]" (Hofstede, 1991; Lincoln and Kalleberg, 1990).

2. Culturally-Driven Normative Systems

In this research, values and practices are viewed as the basic building blocks of culture (Hofstede, 1991; House et al, 2004), hence, culturally-driven normative systems. This research characterizes culturally-driven normative systems of Japanese and American teams along value-practice dimensions through observations and literature survey. Specifically, we conducted four case studies using the ethnographic approach (Spradley, 1979) between April and August 2003. All four projects were joint-venture projects between Japanese and American firms located near the San Francisco Bay Area. Thus, we had a good control over the broader legal and political regulative institutional context (Scott, 2001).

Cultural Values: Hofstede (1991) defines values as conceptions of the preferences and feelings in certain states of affairs with an arrow to a plus or a minus side. Cultural values can be seen as the driver of preferred or desirable behaviors, when participants make decisions or coordinate with each other. We call the behavior, "micro-level behavior" (Jin and Levitt, 1996), which can be observed by focusing on how participants make decisions and communicate with others. Therefore, this

[1] His proposition implies that members of a given cultural group will show better performance when working within their preferred organization structure.

research extends the term "*cultural values*" to refer to the preferences people use to make work-related and communication-related decisions in projects. For instance, based on our observations, Japanese workers tend to seek consensus before making decisions, while Americans prefer to decide independently. We observed that Japanese and American workers have distinctly different patterns of micro-level behavior. These observations are consistent with existing literature (Nakane 1970). In addition, *value differences* are linked to national culture (Hofstede, 1991). Hofstede's work[2] provides a useful set of dimensions against which value differences can be measured. For instance, the *individualism-vs.-collectivism* index Hofstede proposes can explain why Japanese people tend to seek consensus among team members, since Japanese workers are high on the Collectivism scale. In collectivist countries, "harmony should always be maintained and direct confrontations avoided" (Hofstede, 1991, p.49-78). Based on our observations, harmony and trust among group members are key aspects of Japanese workplace culture, and can be seen in many different activities, including meetings and contracts. Thus, lower individualism, high collectivism countries like Japan tend to have group-based decision-making.

Cultural Practices: Scott (2001) asserts that norms specify conceptions of appropriate business practices to define legitimate means and processes to pursue valued ends. Thus, this research extends the meaning of "*cultural practices*" to include norms that regularize specific project management styles and organization structures. Based on observations, *practice differences* at the project team level are characterized by three organizational elements: the level of centralization of authority, the level of formalization of communication, and the depth of the organizational hierarchy. Different cultures in different countries tend to set these organizational elements differently, because different norms prescribe different reasoning and legitimacy for each of these organizational elements. Our ethnographies found that Japanese project teams tend to have multiple levels of hierarchy and to be more centralized, while American firms usually adopt a flat organization hierarchy and decentralized authority. These observations are consistent with existing literature (Lincoln & Kalleberg, 1990) (see Table 1).

Table 1 summarizes the two culture dimensions (*cultural value* and *cultural practice*), their attributes, and the values of these attributes for Japanese and American cultures. At project level, each nation has its own sets of micro-level behavior and organizational style, comprising culturally driven normative systems.

Table 1: **Summary of Culturally Driven Normative Systems**

Cultural values	**Culture A (American)**	**Culture J (Japanese)**
Decision making	Individual decision making	Consensual decision making
Communication	Individually-based	Group-based

Cultural practices	**Culture A (American)**	**Culture J (Japanese)**
Centralization	Decentralized authority	Centralized authority
Formalization	Medium level of formalization	High level of formalization
Organizational hierarchy	Flat level of hierarchy	Multiple levels of hierarchy

3. Computational Simulation Model

The heterogeneous normative multi-agent simulation model of this research is developed based on the Virtual Design Team (VDT) model. The VDT[3] model is adopted as a virtual organization laboratory for three reasons: 1) the VDT model was built to design project organizations, the same unit of analysis as this research, 2) the large numbers of organizational and individual level behavioral parameters available in the VDT model can potentially represent culturally-driven normative systems with some fidelity, and 3) the VDT model has been validated by many previous researchers (e.g., Thomsen et al, 1999). Furthermore, the VDT model fulfills the three key criteria for being used as a "theorem prover" (Burton & Obel, 1995) - reality, content, and structure - to examine hypotheses. Therefore, this research uses the VDT model to analyze the effects of organizational and individual normative differences.

The VDT model (Jin and Levitt, 1996) succeeded in extending the information processing view (March and Simon, 1958; Galbraith, 1973, 1977) by measuring the fit between information processing capacity and information processing demand at the level of an individual actor, called a "neo-information processing view" (Burton and Obel, 2004). In this view, this research encodes stochastic patterns of individual actors' behaviors in decision making and communication driven by differing *cultural values*, based on observations and a literature survey. In other words, we set heterogeneous types of agents in the VDT model. Similarly, this research models organization structures and stochastic decision-distribution patterns driven by differing *cultural practices*. In addition, task complexity and team experience are set as idealized context variables. We assume two independent variables reflect the effects of changes in values and practices: *micro-level behavior of*

[2] Hofstede proposed using four dimensions to describe cultural differences among 53 countries including Japan and the United States: 1) power distance, 2) individualism vs. collectivism, 3) masculinity vs. femininity, 4) uncertainty avoidance, and 5) long term orientation vs. short term orientation.

[3] We use SimVision®, educational version 3.11.1, which was developed by Vité Corporation and is licensed from ePM, LLC, Austin Texas. Please see the website for more information: < http://www.epm.cc/ >

actors (cultural values), *organization style* (cultural practices), over the full range of our context variables of *task complexity*, and *team experience* (Figure 1).

- The *micro-level behavior* of actors is related to their *cultural values*, and refers to actors' decisions about how to handle exceptions and how to communicate with others. Since *cultural val*ues form the basis of how people behave and how they make decisions, *cultural values* are linked to micro-level behavior in the VDT model. We assume that the American behavior pattern is the same as the original set of micro-behavior parameters in the VDT model, because the VDT model was developed and calibrated in American firms (Christensen, 1993; Thomsen, 1999). We create a Japanese behavior pattern by manipulating two sets of micro-behavior parameters that are related to decision-making and communication behaviors respectively, based on our observations and the extant literature (Hofstede, 1991; Lincoln and Kalleberg, 1990; Aoki, 1992). We set two types of micro-behavior patterns to represent Japanese and American styles (Figure 1-1)
- *Organization style*, which is linked to *cultural practices,* refers to the organizational parameters within the VDT model that determine the exception handling paths and authority levels of decision makers. Since cultural norms within an organization specify appropriate and legitimate means and processes (Scott, 2001) that enable the organization to conduct a project, *cultural practices* are linked to an organization's structure style. Specifically, we set three organizational parameters based on our observations: the centralization level, formalization level, and depth of organizational hierarchy. This set of three organizational parameters represents each nation's typical organization style. For instance, the American organization style is set to a low level of centralization (i.e., more decentralized), a medium level of formalization, and includes direct supervision links between the project manager and subordinates. In our analysis, we set two types of typical organization styles to represent the Japanese and American styles (see Figure 1-2).
- In building a model that predicts project performance, we consider one aspect of contingency theory (Galbraith, 1977; Thompson, 1967) to define context: *task complexity.* We examine four different levels of task interdependencies: pooled, sequential, reciprocal, and intensive workflows (Thompson, 1967; Bells and Kozlowski, 2002). These dependencies represent a scale of task complexities, from lowest to highest, respectively (see Figure 1-3).
- The level of team experience is also taken into consideration as a second context variable in order to explore the effects of team mutuality on team performance (see Figure 1-4). Team mutuality indicates that a project team has had a previous experience working together.

4. Design of Virtual Experiments

The main purposes of the intellective experiments are as follows:

--: To study the effects of changes in micro-level behavior patterns (*cultural values*)

--: To study the effects of changes in organization structure styles (*cultural practices*)

--: To study the relationships between micro-level behavior patterns (*cultural values*) and organization structure styles (*cultural practices*) for the full range of possible task complexity and team mutuality contexts.

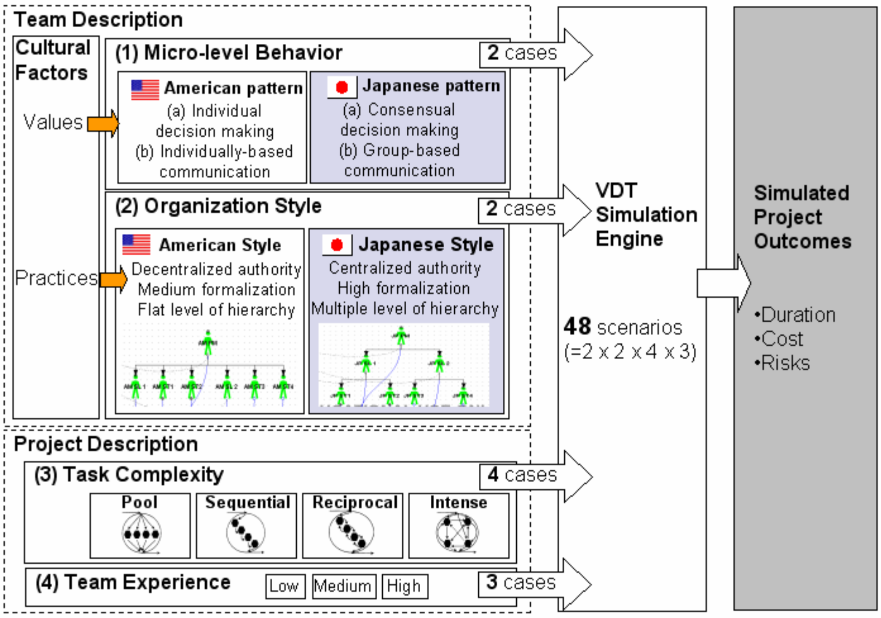


Figure 1: **Modeling Framework**

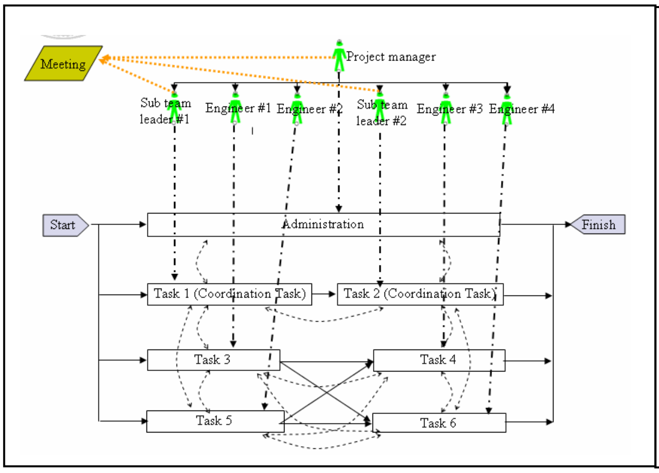


Figure 2: Example of American Organization Structure Type with Intense Complexity

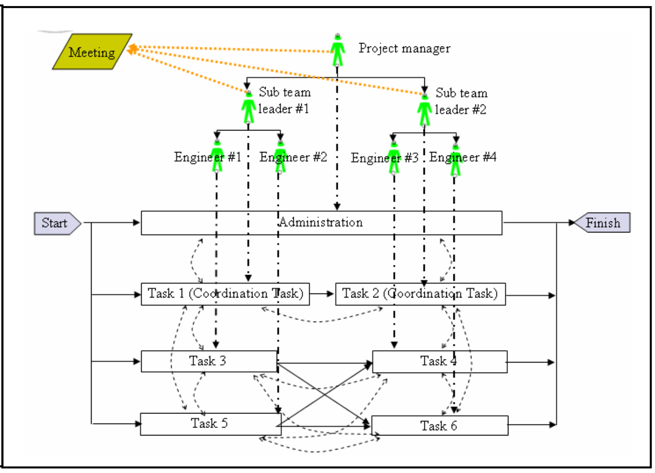


Figure 3: Example of Japanese Organization Structure Type with Intense Complexity

Note: These figures illustrate examples of the *intense* coordination complexity cases. As shown in Figures 2 and 3, both organizations have the exactly same workflow and required work volume as the intense complexity cases. All teams are composed of seven members, including one project manager, two sub-team leaders, and four sub-team members. We change only either structure types or micro-level behavior patters actors possess.

——▶ shows precedence links among tasks

·········▶ shows rework and communication links among tasks

-·-·▶ shows work assignment between team members and tasks

For experimental purposes, the actor and task configurations are identical[4] (Figures 2 and 3 show examples of the *intense* coordination complexity cases with higher numbers of interdependence links between tasks). As shown in Figure 1, we simulated a total of 48 scenarios (2 organization styles x 2 micro-behavior patterns x 4 task complexity levels x 3 team situation levels).

5. Results: Analysis of the Effects of Culturally-Driven Normative Systems on Team Performance

The VDT model, a multi-agent based simulation model, is designed to predict duration, cost, and two kinds of process quality risks as measures of team performance, as shown in Table 3. At first glance, there is no significant difference in the project duration between Japanese and American structural styles. However, differences appear in the hidden work volume, cost, and project quality risks. The amount of hidden work volume is a good proxy for both project duration and cost (Levitt and Kunz, 2002), since the amount of direct work remains constant for all scenarios. Even if duration is apparently the same, hidden work volume presents potential risks of increased cost and duration, as they cause non-critical path tasks to take longer, and thus reduce overall project slack. We analyzed three dependent variables, 1) hidden work volume, 2) product quality risks (see Note 3), and 3) project quality risks (see Note 4) , to measure the impact of changes in elements of culturally-driven normative systems (organization styles and micro-level behaviors) on team performance.

Based on the cultural model described above, we carried out an analysis of the impact of cultural factors on relationships between organization style, team cultural behavior patterns, task complexity, and team experience.

Figure 4 and 5 illustrate the effects of organization structure styles on process quality metrics. The hidden work volume increases as level of task complexity increases. This implies that the idealized case can appropriately capture a basic proposition of contingency theory: “the greater the uncertainty of the task, the greater the amount of information that has to be processed between decision makers during the execution of the task.” (Galbraith, 1974) In the cases of medium task complexity, the American style has less hidden work volume[5] than the Japanese organization style. On the other hand, in the case of high task complexity, this tendency reverses. In particular, when team experience is low, the American style has less tolerance for high task complexity than does the Japanese style.

Figures 6 and 7 show the effects of changes in micro-level behavior patterns on hidden work volume. The effect of changes in micro-level behavior patterns is smaller than the effect of organization style. However, organizational performance of workers who have the culture’s preferred micro-level behavior is positively correlated to the use of each culture’s typical organization style, in cases of medium to high task complexity. In the case of pooled and sequential workflow, the differences between Japanese and

[4] Actor and task configurations include the actors’ skills, the skills required by tasks, the duration of tasks, the hourly salary of actors, the task responsibility assignment, and the total number of team participants. All teams are composed of seven members, including one project manager, two sub-team leaders, and four sub-team members.

[5] Less hidden work volume implies better performance

American behavior patterns are relatively small. This implies that increasing task complexity amplifies the impact of *cultural values vs. cultural practices* mismatches, as we would expect, since it increases the frequency of exceptions that will arise in executing direct tasks (Galbraith 1973).

As shown in Figures 8 and 9, there are no significant differences between the Japanese and American styles in terms of predicted product quality (component quality) risks. However, the Japanese organization style tends to have significantly lower project quality (system integration) risks than the American organization style. The more centralized Japanese structure and close supervision by first level managers with lower spans of control in the deeper Japanese hierarchy imposes tight control on information exchange and exception handling for both Japanese and American workers. So this prediction has good face validity.

When comparing relative magnitude of changes in organization style and behavior patterns, changes in organization style have a larger impact on hidden work volume than changes in behavior patterns.

Table 2: **Summary of Simulated Results**

	Task Complexity **Low** ⟵⟶ **High**							
	Pooled		**Sequential**		**Reciprocal**		**Intense**	
Structural Style	**Type *J***	**Type *A***	**Type *J***	**Type *A***	**Type *J***	**Type *A***	**Type *J***	**Type *A***
Duration (Critical Path Method)								
Duration (Months)	8.0	8.1	29.6	28.8	30.7	29.6	13.5	13.1
Standard deviation	(0.18)	(0.15)	(1.30)	(0.80)	(1.60)	(1.00)	(1.10)	(0.60)
Comparison	Type *J* = Type *A*		Type *J* = Type *A*		Type *J* = Type *A*		Type *J* = Type *A*	
Hidden Work Volume								
Hidden Work Volume (Person-months)	3.54	4.46	14.60	12.57	26.02	21.13	29.19	38.15
Comparison	Type *J* < Type *A*		Type *J* > Type *A*		Type *J* > Type *A*		Type *J* < Type *A*	
Cost								
Cost ($1,000)	281	288	355	343	431	401	446	497
Standard deviation	(2.65)	(2.78)	(27.49)	(17.16)	(47.60)	(33.89)	(33.41)	(56.28)
Comparison	Type *J* < Type *A*		Type *J* > Type *A*		Type *J* > Type *A*		Type *J* < Type *A*	
Functional (Product) Quality Risks								
Product Quality Risk Index	0.469	0.468	0.466	0.464	0.467	0.461	0.478	0.480
Standard deviation	(0.044)	(0.037)	(0.037)	(0.041)	(0.035)	(0.034)	(0.033)	(0.022)
Comparison	Type *J* = Type *A*		Type *J* = Type *A*		Type *J* = Type *A*		Type *J* = Type *A*	
Project Quality Risk								
Project Quality Risk Index	-[6]	-	0.267	0.437	0.284	0.467	0.279	0.472
Standard deviation	-	-	(0.044)	(0.067)	(0.037)	(0.046)	(0.031)	(0.033)
Comparison	-		Type *J* < Type *A*		Type *J* < Type *A*		Type *J* < Type *A*	

Note:

(1) Total simulated work volume is the sum of production work volume and coordination work volume (Jin and Levitt, 1996, pp175)

Hidden Work Volume = Total Simulated Work Volume – Designed Work Volume

(2) For each scenario, we run 100 trials and calculate means and standard deviations.

(3) Product quality risk represents the likelihood that components produced by the project have defects based on rework and exception handling (Jin and Levitt 1996, pp179)

(4): Project quality represents the likelihood that the components produced by the project will not be integrated at the end of the project, or that the integration will have defects based on rework and exception handling (Jin and Levitt, 1996, pp179).

[6] Since there are no communication or rework relationships between tasks in the context of pooled workflow, project quality risk is always zero, and so is not shown for those scenarios.

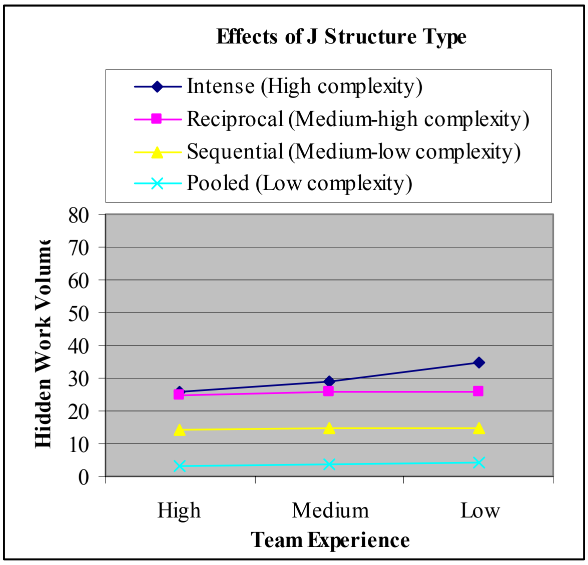


Figure 4: **Effects of Japanese Organizational Structure Type**

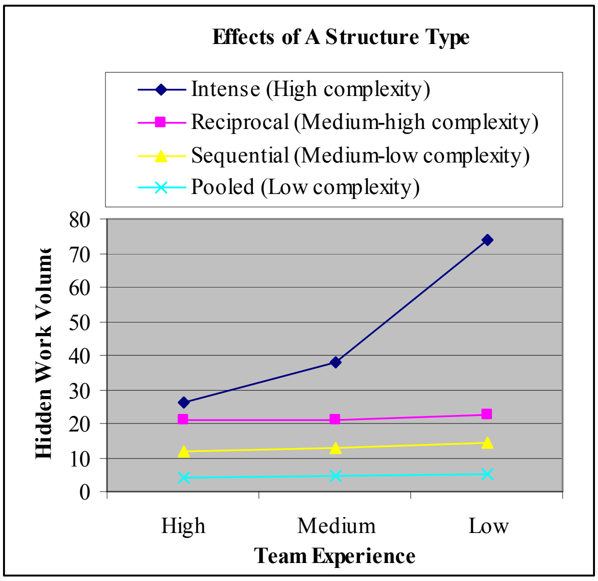


Figure 5: **Effects of American Organizational Structure Type**

Note: This figure compares the performance of Japanese *vs.* American organization structure types. The X axis shows the level of team experience. The Y axis shows total hidden work volume in person-months. Task interdependencies such as pooled, sequential, reciprocal, and intense workflow represent a range from low to high task complexity respectively.

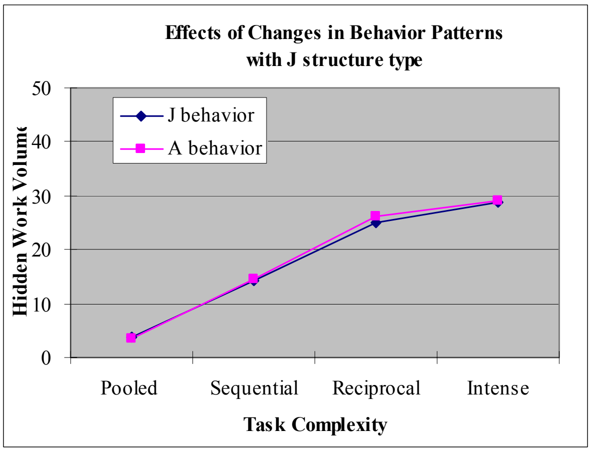


Figure 6: **Effects of American *vs.* Japanese Micro-Level Behavior Patterns with Japanese Organizational Structure Type**

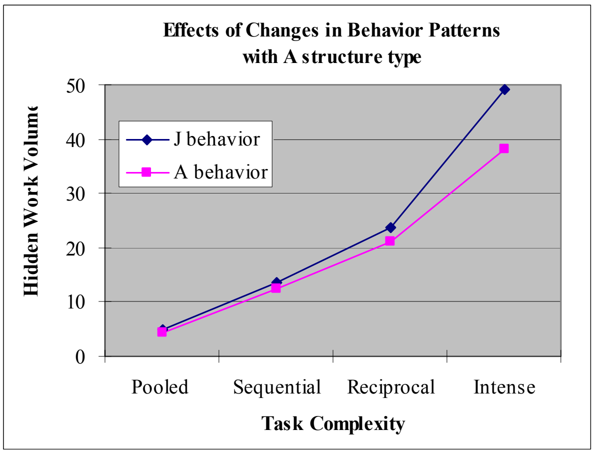


Figure 7: **Effects of American *vs.* Japanese Micro-Level Behavior Patterns with American Organizational Structure Type**

Note: This compares the performance of Japanese *vs.* American micro-level behavior patterns for each structure type. The X axis shows the level of task workflow such as pooled, sequential, reciprocal, and intense interdependencies. Each workflow represents from low to high task complexity respectively. The Y axis represents total hidden work volume in person-months.

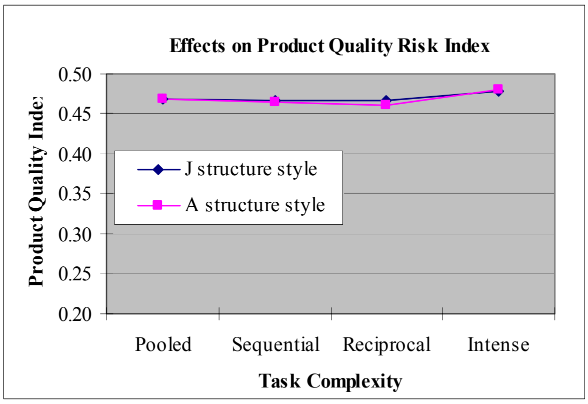


Figure 8: **Effects of Organizational Structure Type on Product Quality Risk**

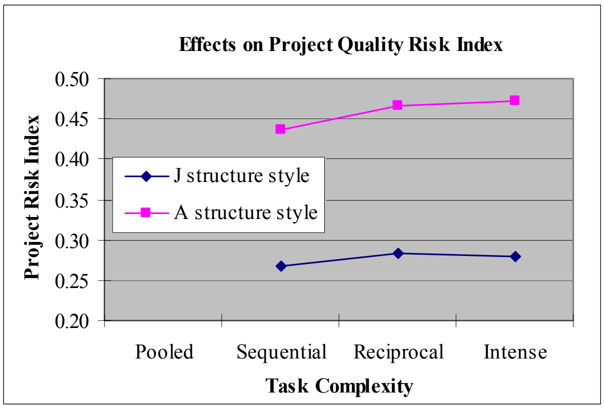


Figure 9: **Effects of Organizational Structure Type on Project Quality Risk**

Note: This compares the performance of Japanese *vs.* American micro-level behavior patterns. The X axis shows the task complexity from pooled (lowest), sequential, reciprocal, and intense (highest) interdependencies. The Y axis represents total hidden work volume in person-months.

6. Discussion

Through computer simulations, we examined the effects of changes in organization structure types (*cultural practice*s) and micro-level behavior patterns (*cultural values*) for a range of possible project situations (task complexity and team mutuality contexts).

Effects of changes in organization structure styles: Each typical organization structure style driven by culture has its own matched project situation in terms of team performance. Specifically, Japanese organization style shows better performance in the case of high task complexity, while American organization style shows better performance in the low and medium task complexity cases. This implies: that managers need to set up appropriate organization styles by considering project situations; that the impact of *cultural practices* is contingent upon the types of task complexity; and that IJV participants need to find equivalent points in *cultural practices* by considering task complexity and environments.

Effects of changes in micro-level behavior: We find support for Hofstede's proposition of preferred coordination mechanisms (Hofstede, 1991), i.e., that team performance is better when management practices are congruent with national cultural values. Hofstede proposes that each culture has a preferred coordination mechanism, implying that workers from each nation show better performance if they use their own preferred management practices (Hofstede, 1991). Our results contribute to the small body of organizational and virtual experimental evidence supporting the importance of congruence between *cultural values* and *cultural practices,* hence, the two normative components. We extrapolate from these findings to conclude: that each *cultural practice* has evolved to match its *cultural values*, in order to maximize efficiency; and that the impact of normative systems (*cultural values* and *cultural practices*) is contingent upon not only types of task complexity, but also types of agents. Inconsistency among normative components can yield undesired results. A project manager must be careful in identifying normative components and in maintaining their consistent relations when designing multi−agent systems (MAS)

Moreover, Hofstede (1991) asserts that each culture's preferred organization style can be predicted from two of his national cultural value indices—power distance and uncertainty avoidance. Figure 10 shows a two-by-two power-distance-uncertainty avoidance matrix, with one of Mintzberg's (1983) five archetypal organizational configurations in each corner, and the fifth, the divisionalized structure archetype, as a kind of "compromise structure type" in the center. Based on case studies, the Japanese and American organization structures are close to the preferred mechanism plotted by Hofstede. Specifically, the Japanese organization structure has relatively high centralization, high formalization, and multiple levels of hierarchy. Hofstede also suggests that Japan is categorized with France as preferring a full bureaucracy, defined as high formalization and well-defined authority hierarchy (e.g., Mintzberg, 1983: Burton and Obel, 2004). Therefore, our experimentation suggests the possibility to predict preferred organization styles from cultural value dimensions proposed by Hofstede (1991).

Relative impacts between organization styles and micro-level behavior patterns: Changes in behavior patterns had less impact on team performance than changes in organization structure. In other words, *cultural practices* have larger impact than *cultural values* on team performance. At this stage, the relative contributions of the organization system or behavior pattern are unknown and cannot be analyzed quantitatively.

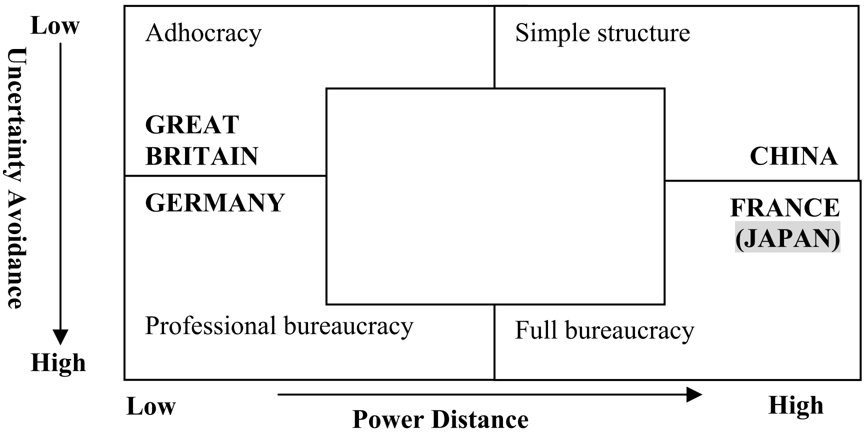


Figure 10: Preferred Coordination Mechanism (Adapted from Hofstede, 1991, p.152)

Note: This figure illustrates the typical organization structure predicted by "power distance index" and "uncertainty avoidance index." "Power distance index" refers to the extent to which the less powerful members of organizations and institutions accept and expect that power is distributed unequally. "Uncertainty avoidance index" indicates the extent to which a culture programs its members to feel comfortable in unstructured situations such as unknown, surprising, and different from the usual. Uncertainty-avoiding cultures try to minimize the possibility of such situations by using strict laws and rules, and safety and security measures

In summary, when organizations assemble joint venture teams with members from different culturally-driven normative systems, a project manager should pay attention to three elements: micro-level behavior (*cultural values*), organization style (*cultural practices*), and project situations (task complexity and team mutuality). Managers need to change their management practice style based on the characteristics and requirements of a given project, because project situations are given at the start of a project, and the micro-level behavior is fixed in the short term (based on national culture), and because the organization style is the only variable a project manager can control. Careless selection of management practices may cause a worst-case scenario in a project. Heterogeneous normative multi-agent models can help managers to find the equivalent point by changing the organization style that provides the best match to their project's characteristics and their team's micro-behavior.

The existing VDT model has known limitations that constrained us in capturing all of the cultural and broader institutional phenomena that emerge in global projects. We were unable to adequately represent factors such as multiple behavior patterns for different workers in a project, additional exceptions caused by work practice differences, organizational learning, and some of the positive impacts—e.g., increased innovation— that might result from cross-cultural interactions. Our experiment focused only on the impact of different patterns of micro-level behaviors and organization structures.

--: Examining the cases where multiple behavior patterns coexist in a project remains an intriguing research focus. We are currently working to extend VDT to permit a modeler to assign different *cultural values* to each "Actor"—i.e., each individual or sub-team—in the project

--: A second constraint was that the current VDT model is not able to parameterize additional exceptions caused by differing values and practices between subgroups of a joint-venture team. In particular, based on our observations, subgroups are likely to have their own standardized low-level work practices, rules and criteria. Our ethnographies provided evidence that such differences generated exceptions between subgroups when selecting standardized criteria for a project, such as those used for safety. Several researchers have addressed differences in institutionalized practices in IJV projects (e.g., Mahalingam et al, 2004).

--: Another VDT-imposed limitation of this work is that we had to assume that team members do not adapt their values or practices during the project. However, researchers have increasingly been interested in how people learn *cultural values* and *cultural practices* from each other (e.g., Orr, 2004).

--: Finally, in the current research, we did not take into consideration potential positive impacts of cultural interactions, through innovation, creativity, and advanced technology. Several researchers have started exploring innovation issues in project-based organizations (e.g., Taylor and Levitt, 2004).

7. Conclusion

Research on IJV projects reveals the difficulties of coordinating cross-cultural teams. Our research sheds light on some effects of the increased internal complexity that IJV project teams face when *cultural values* and *cultural practices* are misaligned. It makes an initial attempt to predict the impacts of differing normative systems on team performance in IJVs through virtual experimentation. We conducted ethnographic interviews to understand and encode *cultural values* and *cultural practices* into the parameters of the VDT model and then characterized the performance outcomes that emerge in global projects involving both Japanese and American cultures, represented along cultural value-practice dimensions.

The effects of changes in micro-level behavior patterns and organizational control styles show interesting correlations between *cultural values* and *cultural practices*, and gives initial evidence that these parameters have been encoded correctly, since our model predictions align with extant theory. These findings not only extend application of the current VDT model to address the case of heterogeneous normative multi-agents, but also demonstrate a possible framework for modeling distinguishing culturally-driven normative factors that emerge in global projects. In addition, our work contributes in a small way to using simulation to bridge the gap between cultural-cognitive psychology as micro-level theory and sociological organization science as macro-level theory.

We have argued earlier (Levitt et al, 1999) that global projects provide an ideal field setting in which to explore the effects of institutional clashes on the behavior and performance outcomes of organizations. Global projects bring together participants from multiple national, organizational and professional cultures. And all projects have unusually clear goals and metrics compared to most other organizational forms; they have a finite start and end date—often with durations that are less than a typical PhD degree—and clearly defined participation. Currently, there are intriguing and unexplored research opportunities to study dynamics of normative systems in inter-cultural, inter-organizational and inter-institutional settings, such as global projects.

Acknowledgement

We thank the Collaboratory for Research on Global Projects (CRGP) <http://crgp.stanford.edu/> and the Virtual Design Team (VDT) research groups at Stanford University for their insightful suggestions and comments on this project.

Notice of Previous Presentation

The preliminary version of this paper was presented at the 2004 North America Association for Computational Social and Organizational Science (NAACSOS) conference at Carnegie Mellon University, and was awarded "best graduate paper" in the Ph.D. Student Research Abstract Competition. This paper focuses more on differing normative systems observed in global projects and developing a heterogeneous normative multi-agent model, providing a different view from the previous paper (Horii et al, 2004).

References

Aoki, M., 1992. Decentralization-Centralization in Japanese Organization: A Duality Principle, *Japanese Political Economy* Vol. 3, Stanford University Press, pp. 142-169

Beamish, P.W. and Delios, A., 1997. Incidence and Propensity of Alliance Formation, *In Cooperative Strategies: European Perspectives*, San Francisco, CA: New Lexington Press

Bell, B.S., and Kozlowski, S.W.J., 2002. A Typology of Virtual Teams: Implications for Effective Leadership, *Group and Organization Management*, Vol. 27 (1), pp. 14-49

Burton, R.M., 2003. Computational Laboratories for Organization Science: Questions, Validity and Docking, *Computational and Mathematical Organization Theory*, Vol. 9, pp. 91-108

Burton, R.M, and Obel, B., 1995. The Validity of Computational Models in Organization Science: From Model Realism to Purpose of the Model, *Computational and Mathematical Organization Theory*, Vol. 1, pp. 57-71

Burton, R.M, and Obel, B., 2004. *Strategic Organizational Diagnosis and Design: The Dynamics of Fit,*" Norwell, MA: Kluwer Academic Publishers, 3rd edition

Christiansen, T., 1993, Modeling Efficiency and Effectiveness of Coordination in Engineering Design Teams: VDT-the Virtual Design Team, *Civil Engineering department*, Stanford, CA: Stanford University

Davis, S.M., 1984. *Managing Corporate Culture*, Cambridge, MA: Ballinger

Dooley, K., 2002. Simulation Research Method, *The Blackwell Companion to Organizations*, Chapter 36, Malden, MA: Blackwell

Galbraith, J.D., 1974 *Organization Design: An Information Processing View*, INTERFACES 4

Galbraith, J.D., 1977 *Organization Design*, New York, NY: Addison-Wesley Inc.

Hofstede, G., 1991. *Cultures and Organizations: Software of the Mind, Intercultural Cooperation and its Importance for Survival*, New York, NY: McGraw-Hill

Horii, T., Jin, Y., and Levitt, R.E., (forth coming), Modeling, and Analyzing Cultural Influence on Project Team Performance, *Computational and Mathematical Organization Theory*

House, R.J., Hanges, P.J., JAvidan M., and Gupta, V., 2004. *Culture, Leadership, and Organizations: the GLOBE Study of 62 Societies*, Thousand Oaks, CA: Sage publications Inc.

Jin, Y., and Levitt. R.E., 1996. The Virtual Design Team: A Computational Model of Project Organizations, *Computational and Mathematical Organization Theory*, Vol. 2(3), pp. 171-196

Levitt, R.E., Cohen, G.P., Kunz, J.C., Nass, C.I., Christiansen, T., and Jin, Y., 1994. The Virtual Design Team: Simulating How Organization Structure and Information Processing Tools Affect Team Performance, in Carley, K. M. and M. J. Prietula, editors, *Computational and Mathematical Organization Theory*, Lawrence Erlbaum, Associates, Publishers, Hillsdale, NJ.

Levitt, R.E., Thomsen, J., Christiansen, T.R., Kunz, J.C., Jin, Y., and Nass, C.I., 1999. Simulating Project Work Processes and Organizations: Toward a Micro-Contingency Theory of Organizational Design, *Management Science* Vol. 45 (11), pp. 1479-1495

Levitt, R.E., and Kunz, J.C., 2002. Design Your Project Organization as Engineers Design Bridges, *CIFE technical paper*, September, Stanford University

Lincoln, J.R., and Kalleberg, A.L., 1990, *Culture, Control, and Commitment: A Study of Work Organization and Work Attitudes in the United States and Japan*, Cambridge, UK: Cambridge University Press

Mahalingam, A., Levitt, R.E, and Scott, W.R., 2004, Cultural Clashes in International Infrastructure Development Projects: Which Cultures Matter? In: *Proceedings of the International Symposium of CIB W92*, CIB W92, Las Vegas, USA

Mintzberg, H., 1983. *Structure in Five: Designing Effective Organizations*, Englewood Cliffs, NJ: Prentice-Hall

Nakane, C., 1970. *Japanese Society*, Berkeley and LA, CA: University of California Press, 1970

Orr, R., 2004, Coping With Cognitive-Cultural, Normative and Regulative Institutional Asymmetry On Global Projects: A learning Perspective, *CIB TG23 Symposium*, Bangkok, Thailand

Ouchi, W.G., 1981. *Theory Z: How American Business can meet the Japanese Challenge*," Reading, MA: Addison-Welsley

Redfield, R., 1948. Introduction to B. Malinowski: *Magic. Science and Religion*, Boston: Beacon press.

Schein, E., 1989. *Organizational Culture and Leadership*, San Francisco, CA: Jossey-Bass, 1985, 2nd ed., 1992

Scott, W. R., 2001. *Institutions and Organizations,"* Thousand Oaks, CA: Sage publications Inc., 2nd Ed.

Sullivan, J.J., and Nonaka, I., 1986. The Application of Organizational Learning Theory to Japanese and American Management, *Journal of International Business Studies*, Vol. 17, No3, pp. 127-147

Taylor, J.E., and Levitt, R.E., 2004. A New Model for Systemic Innovation Diffusion in Project-Based Industries, *Project Management Institute International Research Conference*, London, England

Thompson, J. D., 1967. *Organization in Action: Social Science Bases in Administrative Theory*, New York, NY: McGraw-Hill

Thomsen, J., 1988. Virtual Team Alliance (VDA): Modeling the effects of Goal Incongruence in Semi-Routine, Fast-paced Project Organizations, *Civil Engineering department*, Stanford, CA: Stanford University

Thomsen, J., Levitt, R. E., Kunz, J., Nass, C., and Fridsma, D., 1999. A Trajectory for Validating Computational Emulation Models of Organizations, *Computational and Mathematical Organization Theory*, Vol. 5(4), pp. 385-401

Increasing Software Infrastructure Dependability through a Law Enforcement Approach

Gustavo Carvalho[*] Rodrigo Paes[*] Ricardo Choren[†] Paulo Alencar[††] Carlos Lucena[*]

[*]Depto de Informática – PUC-Rio
Rua Marquês de São Vicente, 225
Rio de Janeiro, Brasil, 22453-900
{guga, rbp, lucena}@inf.puc-rio.br

[†]DE9 - IME
Praça General Tibúrcio 80,
Rio de Janeiro, Brasil, 22290-270
choren@de9.ime.eb.br

[††] Computer Systems Group
University of Waterloo
Waterloo, Ontario, N2L 3G1 Canada
palencar@csg.uwaterloo.ca

Abstract

Software systems are increasingly becoming distributed, open and ubiquitous assets. While open system components are often autonomous, they behave unpredictably when unforeseen situations arise. Taming this uncertainty is a key issue for dependable open software development. This work proposes a law enforcement approach that uses risk analysis to develop dependable open systems. We present law enforcement as a suitable technique to deal with dependability requirements in open systems. Laws impose execution rules and limits, creating a boundary of tolerated component autonomy and fostering the development of trusted systems. We also show that risk analysis methods can help the assessment of dependability alternatives. The approach models dependability requirements as risks that guide the specification of laws. Laws play an essential role in the system architecture and implementation developed to enforce the laws and support the approach.

1 Introduction

Recently, there is widespread interest in software technology and in its associated advances as they affect and stimulate the global economy. Software permeates every aspect of our lives, and is increasingly becoming distributed, open and ubiquitous assets.

Openness has led to software systems that have no centralized control and that are composed of autonomous entities (Agha, 1997). These entities may enter and leave the environment at their will, and they may even have conflicting interests (Fredriksson et al, 2003). Multi-agent auction systems are examples of such open and distributed applications (TAC, 2004; Zambonelli et al., 2003).

Further, open systems need to rely on critical infrastructures that constitute the backbone for the delivery of their essential services. This is not only because open systems are more prone for overloads, attacks or failures, but also because they need to deal with uncertainty (Neumann, 1995). While open system components are often autonomous, they behave unpredictably when unforeseen situations arise. Taming this uncertainty is a key issue for dependable open software development.

Law enforcement can be used as a suitable technique to deal with dependability requirements in open systems. Laws define interfaces for the components that can be present and interact in an open system. This interface imposes execution rules and limits, creating a boundary of tolerated autonomous behavior and fostering the development of trusted systems.

However, a high level of dependability usually has a side effect: low performance. This problem happens because to build dependable software developers generally include extra, often redundant, code to perform the necessary checking for exceptional states, and recovery from faults (Sommerville, 2004). Moreover, these additional design, implementation and validation efforts increase significantly the development costs.

On the other hand, risk analysis methods can help the assessment of dependability alternatives. They can be used as a criterion to establish an order of relevance or importance in dependable software development. We believe that risks can offer a structured method to specify, develop, monitor and maintain system requirements and to foster dependability.

This work proposes a law enforcement approach that uses risk analysis to develop dependable open systems. This approach models dependability requirements as risks. These risks guide the specification of laws that play an essential role in the system architecture and implementation developed to enforce the laws and support the approach. The Figure 1 shows an overview of our approach.

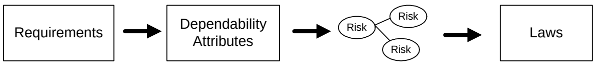


Figure 1 - The Law Enforcement Approach

The contribution of this paper is threefold. First, we propose using a risk based approach to deal with dependability. Second, we propose a conceptual framework to represent laws. At last, we propose mixing risk-based and law enforcement approaches to meet dependability attribute requirements.

The organization of this paper is as follows. In section 2, we discuss how risks can be used to structure dependability attributes. In section 3, we describe the law enforcement approach and the proposed architecture. Section 4 shows how risks are related to laws. In section 5, we show a case study that illustrates the approach. Related work is described in Section 6. Finally, our conclusions are in Section 7.

2 Structuring Dependability Attributes as Risks

Systematic exposition of the concepts of dependability consists of three parts: the threats to, the attributes of, and the means by which the dependability is attained (Avizienis et al., 2001). Besides threats, we aim to include beneficial consequences or opportunities that could also arise from system execution and are described as emergent behavior.

A dependability attribute can be modeled as risks, represented as a sequence, or a chain, of cause and consequence states (Figure 2). The chain of states helps to understand the sequence, tracing back to their origin. By this arrangement, it is possible to infer and analyze how we can gauge and alter the consequences, understanding their causes and the links between intermediary states.

We propose that this exposition should include risks to provide a structured method to specify, develop, monitor and maintain systems requirements and the existent challenges, opportunities, threats and limitations identified through the development of the solution. Briefly, this analysis is based on the identification, control and assessment of relationships between cause and consequence states, events and their characteristics.

For example, a cause of a bad consequence in a system is a point at which a failure or an attack can render the system incapable of continuing to satisfy its requirements for dependability attributes. As in a chain, in which a loss of a single link can destroy it entirely, a specific cause of failure can represent the potential for a disaster triggered at a single point (Neumann, 1995).

Non-functional dependability requirements derive new functional ones that specify how results may be avoided, given incentives or tolerated. In the proposed approach, functional and non-functional requirements could be expressed as laws and consequences that help the achievement of the objectives.

The dependability specification process consists of some activities, including the identification of cause types; their classification; the mapping of classes to metrics; and the identification of functional dependability requirements that reduce the probability of bad consequences or even give incentives for the occurrence of good results.

Nevertheless, it may become impossible to gather all the causes and consequences a priori and, therefore the risk assessment process begins after a relevant group of items has been identified. The assessment is performed considering the severity of each consequence, the probability that it will arise, and the probability that a result will be originated from it. For each item, the outcome of the risk assessment process is a statement of acceptability.

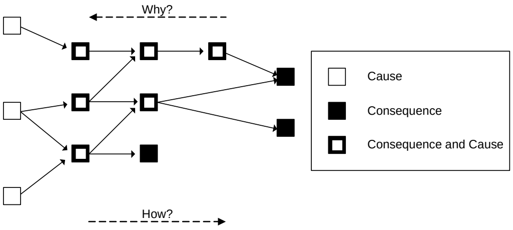


Figure 2 - Chain of cause and consequences instance

Methods, used to understand and analyze the chain of causes and consequences consider both directions of a chain (Figure 2). Analyzing the chain of consequences by looking in the direction of the past, the method tries to find out the origins of the problems or opportunities. This approach is concerned about answering why the consequence can happen. On the other hand, another approach is concerned with answering how some causes would influence the system behavior and the dependable attributes. This analysis begins by identifying the root causes and tries to derive the consequences through the evaluations of future states in the chain.

Some of the most important dependability attributes are reliability, safety, security and survivability. Next, we present a way to model these attributes as risks and the consequences of this approach.

2.1 Reliability

The chain of causes and consequences of reliability can be described as follows (Figure 1). Human errors and mistakes are the primary cause of faults. Due to a lack of understanding of what exactly is happening, sometimes the same fault can result in a normal behavior or in a system error, depending on the event that generates the transition. Likewise, errors can lead to failures only when the fault code is executed with inputs which expose the software failure. Therefore, the reliability can be measured as the probability that the input will cause erroneous outputs.

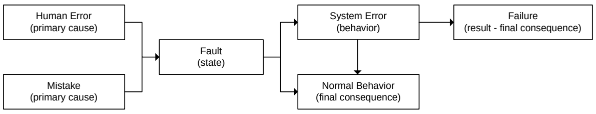


Figure 1 – Chain of causes and consequences of reliability issues

Table 1 shows an example of an availability risk, its consequences, the decisions made to deal with these consequences, and the goals that need to be achieved to implement the decisions.

Table 1 - Availability specification scenario

Risk Identification	Actor unavailability at the beginning of the process.
Risk Description	There may not be available actors in the system for the initial interaction between the participants. There are no available sellers at the moment which a buyer decides to start a negotiation.
Consequences	The system looses its purpose
Decision	For buyer agents: advertisement and promotions. For seller agents: good client base, minimize costs. Monitor and enhance system capability
Decision Goal	Maintain the system working.

2.2 Safety

To assure safety, you must ensure that accidents do not occur or that the consequence of an accident is minimal (Figure 2). Accidents are caused by hazards and have damages as consequences.

Table 2 - Safety specification scenario

Risk Identification	Risk of lack payment by buyer agents
Risk Description	Buyer agent must pay for the acquired ticket. Payment may not happen.
Consequences	Non-payment implies in financial losses for sellers
Decision	If an agent does not pay, the agent will have the ticket cancelled and it will not be allowed to participate in future negotiations, unless… Each buyer agent has a negotiation insurance. If an agent does not pay, besides having the ticket cancelled, the company may be refunded. A mediator (human/software) may be used to solve conflicts.
Decision Goal	Provide a way to incentive and guarantee the air companies. Avoid bad payer agents to keep closing deals which hampers the system

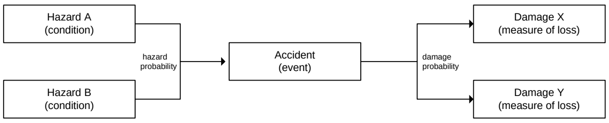


Figure 2 – Chain of causes and consequences of safety issues

Table 2 shows an example of a safety risk.

2.3 Security

Errors can lead to security loopholes, and damages can be caused exploiting this weakness (Figure 3). Some examples of different types of damages include denial of services (unavailable); corruption of programs or data (unreliable, unsafe and unavailable); and disclosure of confidential information (unreliable, unsafe and unavailable).

The chain of causes and consequences of this dependability attribute can be described as follows (Figure 3). Having control of a situation reduces vulnerability. Threats (circumstances) contribute to vulnerability. Vulnerabilities may lead to exposure (possibility) and attacks (exploitation).

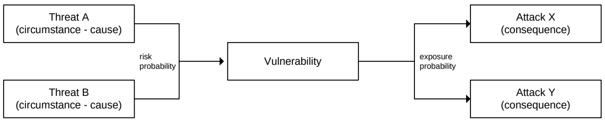


Figure 3 – Chain of causes and consequences of security issues

Table 3 - Security specification scenario

Risk Identification	Disguise – Fake identity – or Agent simulation
Risk Description	The interaction between agents requires the participants' identification. An agent may pose as another, taking advantage of this situation.
Consequences	Less system reliability Losses for the participants Financial losses and lack of trust in the application User rejection as a consequence
Decision	Apply user identification and credentials verification policies on every interaction. To interact, software agents will need an identification in the system and they will also have to guarantee the ID's authenticity.
Decision Goal	Provide a reliable user identification mechanism. Increase the confidence on the systems interactions

2.4 Sustainability & Survivability

A typical software system characteristic is that the properties and the behavior of its components are inextricably intermingled. The successful execution of each system component often depends on the execution of other related components. So, open system approaches must consider this characteristic and its influence to preserve the sustainability of other entities (Fredriksson et al., 2003).

Sustainability is the dependability property concerned with causing or allowing the open system to continue executing for a determined period (Fredriksson & Gustavsson, 2002). Its objective is to help the preservation or to guarantee a harmonious execution environment condition, that is, the behavior of one participant or a group of participants might not prejudice the possibility of other distributed software components are executing in a specific environment.

Sustainability is achieved by considering many other dependability attributes. For instance, to maintain a system sustainable it is desirable to keep it safe, reliable and secure or to provide means to avoid the occurrence or repetition of bad behaviors.

The scenario used to exemplify this attribute specification can be seen as the composition of Table 1, Table 2, and Table 3.

Survivability is concerned with continuing to deliver the service while the system is under attack or even while part of the system is disabled for security and availability issues (Sommerville, 2004). There are many ways to improve the survivability of systems including resistance to attack, attack recognition and recovery from damage techniques. Survivability is directly related with sustainability in open system context.

3 Law Enforcement Approach

In this work, we aim to translate the qualitative or even quantitative criteria presented by dependability attribute to a mechanism that will enforce behavioral rules. Rules are specified as laws and norms. Laws and norms are associated with well-established consequences that are subjected to any participant of an open system. These specifications aim to preserve the dependability attributes.

Laws are identified through a risk-driven analysis of the open system environment. The law enforcement mechanism allows control of the failures and benefits, and it also contributes to tame the uncertainty presented by open systems. This mechanism intercepts some interactions among distributed software components to control and audit the execution flow of conversations.

The enforcement approach aims to contribute to transition from unstable and very unpredictable open systems to the development of dependable, more stable and less unpredictable systems. A risk-based approach can facilitate the understanding of the complexity and the multiple variables involved in open systems development.

3.1 Laws Conceptual Model

In order to obtain a common understanding about laws, we propose a conceptual model. This model is intended for describing the elements that compose a law specification.

A law specification is composed of a description of interaction protocols, norms and time restrictions. These three elements are interrelated in a way that it is possible to specify interaction protocols using time restrictions, norms to control interaction protocols, or even create time sensitive norms.

Interaction protocols define the valid interactions that distributed software components can have and the context where the information exchanged must be interpreted. The specification of interaction protocols using laws allows the protocol enforcement and helps to acquire a better understanding about the problem.

Norms capture the behaviors that are allowable, forbidden, or obligatory. As we mentioned before, norms and protocols work together, complementing each other.

Digital clocks represent the time restrictions and they can be used with protocols or norms as well. Clocks could indicate that a certain period has elapsed producing clock-ticks events.

We use a state machine based approach to specify protocols. A protocol is composed of transitions. Transitions structure the change from previous state to next state, which define what actors can join in a specific part of the conversation, and which are the valid actions. Protocol transitions are activated by sending and receiving messages to software components. However, many other kinds of events can also activate or deactivate transitions. Examples of events are clock-ticks, arrival and sending of messages.

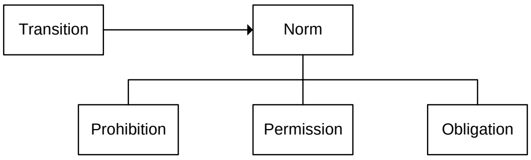


Figure 4 - Transition and norms

Norms prescribe how the distributed software components ought to behave, and specify how they are permitted to behave and what their rights are (Jones & Sergot, 1993). A norm is composed of obligations, permissions and prohibitions. An obligation defines the consequences that the distributed software component (DSC) actions within protocols will have in the future. For instance, the winner of an auction is obligated to pay the committed value. Permissions define the rights of a DSC in a given moment, e.g. the winner of an auction has permission to interact with a bank provider through a payment protocol. Finally, prohibitions define forbidden actions of a DSC in a given moment.

Each norm element (obligations, permissions and prohibitions) has an activation or deactivation condition and consequence. The conditions of activation and deactivation are logical expressions that are evaluated as true or false. The consequence is an instance of an obligation, permission or prohibition. The instance can carry information about the context where the instance was generated. A context is a set of values representing the past DSC interactions, the set of obligations, permissions and prohibitions, and any other value regarding the system execution.

3.2 Law Enforcement Architecture

We propose a law enforcement architecture to guarantee that the specifications will be obeyed (Figure 5) and we developed an infrastructure which includes some communication components that will be provided to DSC developers. This architecture is based on a mediator that intercepts every message and interprets the laws previously described. The main goal of this phase is to provide the infrastructure for the mediation of conversations between components. This phase is also responsible for developing the basic communication components and the interoperability concerns.

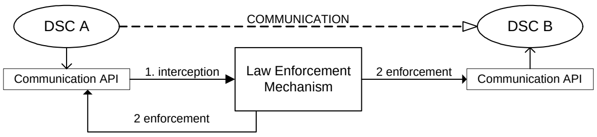


Figure 5 - Law Enforcement Architecture

Depending on the solution domain, it could be necessary to extend this basic infrastructure to attend system requirements. To develop dependable software, this infrastructure must implement some dependability techniques like fault tolerance, error handling and redundancy.

Distributed software components are independently implemented, i.e., this development can be done without a centralized control. The developers may have an a priori access to every specification, protocol descriptions or laws generated during the specification of the open system.

4 Using Laws to Mitigate Risks

In this section, we propose that the risks previously identified must be mitigated using the law enforcement approach and its related concepts. A software engineer has some alternatives or strategies to deal with risks. These strategies deal with the events that could cause the risk and the moment when actions are taken. For example, we can let the event that could cause the risk happen, ignoring its consequences; we can alternatively avoid its occurrence by controlling the process before the execution of determined action, and finally, we can take an action after the occurrence of risks.

Some techniques could be used in conjunction with law enforcement approach to deal with risks (Laprie et al., 1995; Sommerville, 2004). Risk avoidance and risk tolerance are techniques used to minimize risk occurrence, to trap events before they result on bad consequences and to provide means to repair possible damages. Associated with these techniques, we can have a forward strategy to recovery damages or a backward strategy to recover from errors.

To illustrate how the law enforcement approach could be used to improve the dependability attributes, let us consider some examples on how a law enforcement controller can be used. Such a controller can be used, for example, as a checking facility that acts like a fault tolerance mechanism to improve reliability. Besides, the law enforcement approach can be seen a mechanism for assertion execution. In this case, it acts as a protection mechanism to ensure that some erroneous behavior are discovered and corrected before system services are affected. For instance, the interception should not allow a distributed software component that has broken a rule to continue interacting with others or it should even avoid the consequences of this action.

Besides, a law enforcement controller could provide sophisticated interlocks to improve safety. It supports control strategies that reduce the amount of time people have to spend in a hazardous environment.

Furthermore, a law enforcement approach should be used to validate and impose security policies established by the open system. It would impose, for example, restrictions on how and which components could interact with each other.

Finally, a law enforcement controller could be used as a mechanism for assuring that the sustainability laws, derived by the analysis of sustainability attributes and other functional requirements, are fulfilled by the open system participants.

5 Case Study

In this section, we present a specific example to illustrate the application of our approach and highlight its main features.

Suppose an airport where flight companies and passengers have an immersive environment for negotiating flight tickets. This environment is immersive in the sense that the goal of this environment is to enhance computer use by making many computers available throughout a physical environment, and also by making them effectively visible to as many users as possible. Users can have access to the airport services using systems like PDAs and mobile phones. Flight companies and passengers are represented by software agents. Software agents are DSC and they can enter or leave the environment at their own will (Zambonelli et al., 2003).

Flight companies offer tickets for commercial flights. The goal of a flight company is to sell the maximum number of tickets, to increase the user satisfaction and to charge them as much as possible. Passengers use palmtops when they arrive at the airport to buy flight tickets. Each passenger has a specific profile that defines his/her preferences concerning the destination, flight types, maximum acceptable ticket cost, and any other characteristics.

Passenger groups can bargain and get discounts when buying more tickets in a same negotiation process. A group should be formed considering common preference attributes of the participants specified in their personal interests, e.g., the same destination, price, comfort or time of departure.

In the whole negotiation process, a specific step exists for creating groups of interests, where the participant personal profiles are combined aiming to inform other participants that they have close interests. Using this information, it is possible to form a group with close preferences and this group can bargain discounts with the sellers.

By understanding the inherent risks in this case study, it is possible to specify interaction protocols and laws that will regulate the multi-agent system. Figure 6 show the steps related to the protocol specification. The user arrives physically at the airport; it tries to form groups to bargain discounts; it sends messages to the flight companies' agents (FCA) to negotiate their values and attributes; it must pay the ticket to the air flight company which causes the FCA to emit the electronic ticket; it can check in using the electronic ticket, and it leaves the environment.

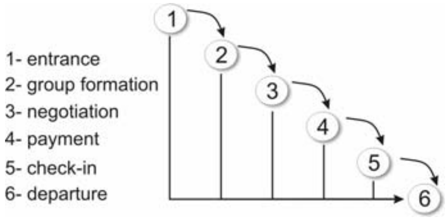


Figure 6 - Agents Execution Scenario

In this paper, we focus on the negotiation phase. The flow of conversation in this phase consists of the following steps:

1- The user agent sends a call-for-proposal message asking for flight tickets. That message contains information about attributes of the flight ticket. These attributes represent the passenger preferences and they are related to, for instance, the preferred departure time or degree of priority on low prices.

2- The flight company offers one ticket flight to the passenger who can, in turn, accept or reject the offer.

3- The user agent accepts the offer and sends a confirmation message.

4- The flight company produces an electronic flight ticket and sends it to the user agent.

5- The user agent receives the ticket and concludes the negotiation.

These steps describe the process, and we have used this description to execute the risk identification and law specification stages.

Table 4 - Risk specification example

Risk Identification	R1
Risk Description	No answers for call-for-proposals messages.
Causes	• C1. Overwork on flight companies' system. • C2. Crashing on flight companies' system.
Consequences	• Co1. The customer agent will be waiting indefinitely. • Co2. The customer could be upset and not use the system anymore. • Co3. If a crash is the cause of the problem, all negotiations between customers and flight companies are unable to continue.
Severity	Medium
Priority	Medium
Dependability attribute	Reliability and sustainability
Design Decision or Rule	We should specify a clock, which will start when the call-for-proposal message is send. After X seconds, the clock activates a norm allowing the customer agent to cancel the negotiation process without penalties.
Rule Goal	• Avoid annoying customers with long waiting times. • Give a specified time for flight companies to execute its market strategies.

We have developed a scenario form that allows us to specify risks, to keep the tracing of causes and consequences, to describe which dependability attributes the consequences affect, and to specify the solution decision. A solution decision could be either a design solution or a law specification.

We use identifiers to represent risks, causes and consequences. In this way, it is possible to construct and identify the chain of causes and consequences related to the inherent risks. Below we show how we use the forms.

The scenario description and the initial risk analysis provide the information needed to specify the protocol. It is important to highlight that, although we are showing the risk identification and protocol specification as sequential steps, they are not. During the protocol specification, it is possible that we discover risks, and during the risk identification, we can discover some protocol elements as well.

Due to space limitations, we do not intend to present a comprehensive case study, but our goal is to highlight the main features of our software dependability approach. Our approach begins with the scenario description and continues through risks and protocol identification. The risk analysis generates new norms, associates clocks, and proposes some protocol modifications. All those concepts compose the laws of the system. Furthermore, we can specify many other laws to deal with the unpredictable behavior of the participant agents. This informal specification should be certainly specified, as one of our next steps, in a formal manner. However, the definition of formal law representations is out of the scope of this paper.

Table 5 - Risk specification example

Risk Identification	R2
Risk Description	Client takes too long to decide if he accepts or rejects the ticket offer. When flight companies offer flight tickets, they reserve a sit for the customer in a way that no other customer can buy that specific sit. However if the client takes too long to answer the request, the flight company could loose the opportunity for sell the ticket to another customer. On the other hand, the client also needs some time to decide whether to accept or reject the offer
Causes	• C3. Undecided customer. • C4. Bad network function. • C5. Bad customer's software function.
Consequences	• Co4. Flight companies could sell the sit to other customers. • Co5. The flight company could be penalized with money losses caused by undecided passengers.
Severity	High
Priority	High
Dependability attribute	Reliability and sustainability
Design Decision or Rule	Specify a clock that after X seconds activates a permission to cancel the negotiation and gives the permission to the flight company. Create a norm forbidding future negotiation participations for the companies that do not follow the time restriction.
Rule Goal	• Give time for customer to decide about acceptance or rejection of offers. • Protect flight companies against undecided customers.

We have implemented the enforcement mechanism following the architecture of the Section 3 and developing a component that extends the Jade (Bellifemine et al., 2001) communication API to provide the redirection of messages. We have also developed an application to monitor the process execution. For example, this application shows the norm activation and deactivation and it collects some metrics that were previously specified.

6 Related Work

Law Governed Interaction (LGI) (Minsky & Ungureanu, 2000) proposes a mechanism to coordinate and control heterogeneous distributed systems. It is based on four basic principles: (1) coordination policies need to be enforced; (2) the enforcement needs to be decentralized; (3) policies need to be formulated explicitly rather than being implicit in the code of the agents involved and they should be enforced by means of a generic, broad spectrum mechanism; (4) and it should be possible to deploy and enforce a policy incrementally. However, this approach does not provide an explicit method to develop and evolve its law enforcement infrastructure.

One of the most interesting related works is the electronic institutions approach (Esteva, 2003; Rodriguez-Aguilar, 2001). In this approach, some concepts related to law enforcement were formalized, software tools were developed to facilitate the institution's design, a textual specification language called ISLANDER was defined, and an infrastructure that mediates agent interactions and enforces the institutional rules on participating agents was developed. Another contribution of this group is the method for rapidly building prototypes of large multi-agent systems using logic programming (Vasconcelos et al., 2004). This method advocates the use of all permitted interactions among the components of the systems. In contrast, the focus of our work is on structuring the development process of open middleware software, thus providing, during this process, development guidelines, structured as risks and dependability attributes.

Cole et al. (2001) proposes a way to identify laws in real world problems. However, his results do not deal with issues related to law enforcement and specification. In addition, Mineau (2003) proposed that laws should be specified using a conceptual graphs approach. This approach supports the validation of rules and uses a very rich and expressive language but it does not propose any enforcement mechanism.

7 Conclusions

Tomorrow's infrastructures will have to face the challenge of survivability: delivering critical services in a timely manner in presence of overloads, attacks and failures (Ellison et al., 1999). This challenge has pushed to the point where proactive risk management is essential. It has become important that software engineers determine whether unwanted events may occur during the development and maintenance of a software system, and make appropriate plans to avoid or minimize the impacts of these events (Neumann, 1995).

In this work, we provide a means for translating the qualitative or even quantitative criteria presented by dependability attributes to a mechanism that will enforce behavioral rules. Rules are specified as laws. Laws are associated with well-established consequences that are subjected to any participant of an open system, and they are identified through a risk driven analysis of the open system environment. This mechanism permits to control the failures and benefits and it contributes to tame the uncertainty presented by open systems.

We believe that this paper represents an advance in the way that dependability attributes requirements can be met and the uncertainty of open environments can be tamed. However, many interesting extensions can be foreseen on the research side, including the formalization of risk specifications with their graphical counterparts, and the development of automatic tracking tools that can support the work of developers and help to improve the productivity of the whole process. Although it is not the focus of this paper, we have considered adding more formalism and methods for describing the laws and automatically generating solutions for verifying it.

Metrics (Fairley, 2002) could be specified in the risk identification phase and the automatic mechanisms provided by the law enforcement infrastructure are able to gather metrics and provide feedback on how efficient these metrics are. In this way, the approach could also help developers to specify system laws and to quantitatively measure their effectiveness.

As a future work, we intend to extend Anote notation (Choren & Lucena, 2004) to represent the design decisions, including the specification of system requirements considering information about risks, interaction protocols, norms activation and deactivation, and any other adaptation that could provide a better understanding of the solution. Furthermore, we also intend to provide a conceptual framework to aid the assessment of existing alternatives of law specification and enforcement mechanisms considering functionality, performance, cost and dependability system properties using for this purpose a risk driven approach.

Acknowledgements

We gratefully acknowledge the financial support provided by the CNPq as part of individual grants and of the ESSMA project (552068/2002-0).}

References

G. A. Agha. Abstracting Interaction Patterns: A Programming Paradigm for Open Distributed Systems, In (Eds) E. Najm and J.-B. Stefani, Formal Methods for Open Object-based Distributed Systems IFIP Transactions, Chapman & Hall, 1997.

A. Avizienis, J.C. Laprie, B. Randell. Fundamental Concepts of Dependability, Re-search Report N01145, LAAS-CNRS, April 2001

F. Bellifemine, A. Poggi, G. Rimassa. “JADE: a FIPA2000 compliant agent devel-opment environment”, Proceedings of the fifth international conference on Autono-mous agents, ACM Press : Montreal, Quebec, Canada, pp. 216-217, 2001.

R. Choren., C. Lucena. Agent-Oriented Modeling Using ANote. Proceedings of the Third International Workshop on Software Engineering for Large-Scale Multi-Agent Systems, SELMAS 2004, Edinburgh, Scotland, May 2004, p. 74-80.

J. Cole, J. Derrick, Z. Milosevic, K. Raymond. Policies in an Enterprise Specifica-tion, In Policies for Distributed Systems and Networks, Springer-Verlag: Lecture Notes in Computer Science, v. 1995,pp.1-17, 2001.

R.J. Ellison, D.A. Fisher, R.C. Linger, H.F. Lipson, T.A. Longstaff, and N.R. Mead. "Survivability: protecting your critical systems", IEEE Internet Computing 3(6), 1999, pp 55-63.

M. Esteva. PhD Thesis, Electronic Institutions: from specification to development, Institut d'Investigació en Intel.ligència Artificial, October 2003, Catalonia – SPAIN.

R. Fairley. Risk-based software estimation. Encyclopedia of Software Engineering. (ed. John J. Marciniak). New York, John Wiley & Sons. Vol 2, 2002, p.1227-1233.

M. Fredriksson et al. (2003) First international workshop on theory and practice of open computational systems. In Proceedings of twelfth international workshop on Enabling technologies: Infrastructure for collaborative enterprises (WETICE), Workshop on Theory and practice of open computational systems (TAPOCS), pp. 355 - 358, IEEE Press.

M. Fredriksson, R. Gustavsson. A methodological perspective on engineering of agent societies. (Eds) A. Omicini and F. Zambonelli and R. Tolksdorf . In Engineer-ing societies in the agents' world, Springer Verlag v. 2203, pp. 10-24, 2002.

A.J.I. Jones, Sergot M. "On the Characterisation of Law and Computer Systems: The Normative Systems Perspective". In Eds J.-J.Ch. Meyer and R.J. Wieringa, Deontic Logic in Computer Science: Normative System Specification, John Wiley and Sons, chapter 12, pp. 275-307, 1993

J.C.Laprie et al. Architectural issues in fault tolerance. In Software Fault Tolerance (M. R. Lyu, ed.) Chichester: John and Sons, pp. 47-80, 1995.

G.W. Mineau. Representing and Enforcing Interaction Protocols in Multi-Agent Systems: an Approach Based on Conceptual Graphs, IEEE/WIC International Conference on Intelligent Agent Technology, 2003.

N.H. Minsky, V. Ungureanu. Law-governed interaction: a coordination and control mechanism for heterogeneous distributed systems, ACM Press, ACM Trans. Softw. Eng. Methodol., v.9, n.3, 2000, pp. 273-305.

P.G. Neumann. "Computer-Related Risks". Published by ACM Press / Addison Wesley. 1995, ISBN 0-201-55805-X, 384pp.

J.A. Rodriguez-Aguilar. On the Design and Construction of Agent-mediated Elec-tronic Institutions. PhDThesis. Institut d'Investigació en Intel.ligència Artificial. 2001, Catalonia - SPAIN.

I. Sommerville. Software Engineering. 7.ed. New York: Addison-Wesley, 2004. 759p.

TAC, Trading Agent Competition, http://www.sics.se/tac/, September 2004.

W. Vasconcelos; D. Robertson, C. Sierra, M. Esteva , J. Sabater, M. Wooldridge. Rapid Prototyping of Large Multi-Agent Systems Through Logic Programming Annals of Mathematics and Artificial Intelligence. August 2004, vol. 41, no. 2-4, pp. 135-169(35).

F. Zambonelli, N. Jennings, M. Wooldridge. Developing multiagent systems: The Gaia methodology, In ACM Trans. Softw. Eng. Methodol., ACM Press, v. 12, n. 3, pp. 317-370, 2003

My Agents Love to Conform: Emotions, Norms, and Social Control in Natural and Artificial Societies

Christian von Scheve
ZiF, University of Bielefeld
Wellenberg 1, D-33615 Bielefeld, Germany
xscheve@informatik.uni-hamburg.de

Daniel Moldt
University of Hamburg, Computer Science Department
Vogt-Koelln-Str. 30, D-22527 Hamburg, Germany
moldt@informatik.uni-hamburg.de

Julia Fix
University of Hamburg, Computer Science Department
Vogt-Koelln-Str. 30, D-22527 Hamburg, Germany
julia.fix@gmx.de

Rolf von Lüde
University of Hamburg, Institute of Sociology
Allende-Platz 1, D-20146 Hamburg, Germany
luede@uni-hamburg.de

Abstract

This contribution investigates the function of emotion in relation to social norms, both in natural and artificial societies. First, the authors briefly illustrate that norms as socially shared mental objects play a crucial role in the dynamics of social structures and social order, in natural societies as well as in artificial systems. Second, the authors address the question how norms are enforced and thereby maintained throughout a social system. In this respect, it is shown that emotions play a crucial role by providing means for intrinsic gratification and sanctioning. The authors consider emotion related sanctions as a cost equivalent to and in many situations perhaps even more efficient than, e.g., resource-driven penalties. Consequently, agents' anticipation of negative emotional outcomes as a consequence of deviant behaviour is supposed to exert social control. Third, the authors outline the possibilities of an application to the socionic multi-agent architecture SONAR

1 Introduction

For some time now apprehensions from the general public as well as from the scientific community have been issued concerning the controllability of artificial intelligence systems, in particular distributed systems based on intelligent autonomous agents. Agent systems are feared to run out of control in such a way that autonomously generated (although probably temporary) goals pursued by a system might contradict (implicit) high-level goals of the designer or user, respectively. Unfortunately, the dilemma arising out of these possible goal conflicts affects some of the core strengths of artificial agent systems: autonomy, flexibility, and discretion. Therefore, means have to be developed that on the one hand ensure the autonomy of the systems in question, and on the other hand avoid conflicts with implicit human high-level goals.

One solution to this problem is the implementation of a system of socially shared norms which is not coerced by the designer, but instead emerges from the mutual interactions of the agents (with actors and/or users). To realise such an approach, it would be beneficial, if not mandatory, to have profound knowledge of and adapt to the computational context the mechanisms of norm emergence, prevalence, and compliance in human social systems (e.g., Dignum et al. 2000; Saam/Harrer 1999).

We argue that *emotions* constitute such a mechanism and that they should therefore be taken into account in the design of agents and especially multi-agent systems (MAS). However, the concept of agents is inspired by and to a large extent also relies on findings from (cognitive) psychology and hence on this discipline's conceptualization of intelligent behaviour, which still is fundamentally based on cognition. Belief-Desire-Intention (BDI) architectures can well be considered an epitome of this perspective on intelligent information processing.

Notwithstanding this, the interrelation of emotion and cognition and the role of emotion in overall intelligent behaviour have been long debated in psychology and have likewise promoted the idea that artificial intelligence (AI) systems could be improved by taking into account mechanisms which are functionally equivalent to emotion in biological systems (Simon 1967; Sloman/Croucher 1981).

At least since Marvin Minsky's programmatic and frequently cited statement that „the question is not whether intelligent machines can have emotions, but whether machines can be intelligent without any emotions" (Minsky 1986: 163), efforts have been increased within the AI community to develop "emotional agents", i.e. software agents capable of utilising mechanisms which are functionally equivalent to emotions in human and non-human animals (for extensive overviews cf. Cañamero (1998), Hatano and associates (2000), as well as Trappl and colleagues (2003)). Unfortunately, until now, the function of emotions in larger societal structures has not been investigated thoroughly, despite some efforts in the area of distributed systems research (e.g., Elliot (1993), Aubé/Senteni (1996), Gmytrasiewicz/Lisetti (2000), and Fix (2004); see also von Scheve/Moldt (2004)).

In view of the fact that distributed (AI) systems are of increasing importance in many areas of application, e.g., electronic marketplaces, automated negotiations, planning and scheduling systems, business process and workflow management, coordination of large-scale open systems, and simulations, it seems reasonable to further investigate the function of emotion in large scale social systems, natural and artificial ones. Pioneering research in the computational study of social norms and emotion has been conducted by Alexander Staller and Paolo Petta (2001); however, this contribution focuses slightly different goals in that it emphasises social structural dynamics.

The article is structured as follows: In the next section we first illustrate the social functions of emotion, both in view of an agent's internal functioning and in view of social interactions. In section three we argue for a model of social control that is fundamentally based on two specific functions of emotion in relation to social norms: the triggering of action incentives (action readiness) and the control of social action. In this model, the emotional commitment to norms in particular ensures actor's compliance with these norms. In the fourth section we then outline how these findings might be applied to the Sonar multi-agent architecture.

2 Social Functions of Emotion

This section examines the functions of emotion in social interactions. The social functions of emotions can conceptually be distinguished from their intra-individual (Levenson 1999), phylogenetic (Turner 2000; Cosmides/Tooby 2000) or ontogenetic functions (Abe/Izard 1999; Holodynski/Friedlmeier 2005).

In order to perform a functional analysis of emotion, we first sketch our definition of emotion and our understanding of functional analysis. Thus, we define emotion as „functional, organised responses to environmental demands that prepare and motivate the person to cope with the adaptational implications of those demands" (Smith/Pope 1992: 36), whereas environmental demands in principal can be both, physical and social. According to this definition, a central function of emotion is the adaptational and beneficial regulation of an agent's behaviour in relation to its environment (Keltner/Gross 1999: 468).

Averill (1992), for example, locates these social functions on three different levels (biological, psychological, and social), Keltner and Haidt (1999) on four levels (individual, dyadic, group, and cultural), Gerhards (1988) likewise on four levels (organism, personality, social structure, and culture) and von Scheve and Moldt (2004) on three levels of analysis (micro, meso, macro). These partitions differ in principle only conceptually and in their ratio of abstraction and/or reduction (cf. also Turner 2002: 30-41).

On the level of an individual agent, emotion performs above all two functions: on the one hand, emotion informs an agent about those events in the social environment that often require immediate, reactive and adaptive behaviour (Schwarz 1990; Clore et al. 1994). For example, annoyance informs about the felt fairness of an action; love informs about degrees of affection and commitment; shame and embarrassment inform about the conformity of an action (cf. Keltner/Haidt 1999). On the other hand, emotions prepare an agent to react adequately upon requirements arising from social situations, e.g. through physiological changes (Clore 1994; Dimberg 1997; Cacioppo et al. 2000). An analogous view is put forward by Oatley and Jenkins (1996: 252), who locate the intraindividual functions of emotion in action readiness and in the structuring of the cognitive system into adequate operational modes.

On the interindividual level, i.e. the expressive and communicative level, there are above all three social functions of emotion that relate in particular to social norms (see Keltner/Haidt 1999):

First of all, emotion expressions allow the attribution of most interactional contingencies, including emotional state, appraisals, intentions, and corresponding interpretations of the situation.

Secondly, emotion expressions may (unconsciously) evoke complementary or reciprocal reactions in context-bound observing actors and therewith contribute to improved bilateral interpretations of a situation. This in turn is a prerequisite for cooperation and the coordination of action.

Thirdly, emotion expressions promote or obstruct specific courses of action and interaction for interacting individuals by exhibiting either motivating or sanctioning clues.

In view of the social environment, within which the agent-environment contingencies and reciprocities emerge, Keltner and Haidt (1999: 507) argue: „Functional explanations refer to the history of some object (e.g. behaviour or trait), as well as the regular consequences that benefit the system in which the object or trait is contained”. We insofar attach our argument to this viewpoint as we focus the *regular consequences* for the system that contains an agent as well as that agent's actions which are in part guided by norm related emotions.

The social functions of emotion in larger social units can be seen at their contributions to identifying social groups and group members (Durkheim 1994), at the ascription of status and power resources (Kemper 1978), at the construction and maintenance of solidarity and cohesion (Lawler et al. 2000), and at the internalisation and retention of social norms, power structures, moral ideas, and ideology concepts (Elster 1999; Hochschild 1979/1983).

Having briefly clarified our position regarding the social functional analysis of emotion, we move towards a more general social scientific analysis of social functions at several levels of abstraction and with regard to common problems of the social sciences and multi-agent system design, a shift in perspective that is also suggested by Castelfranchi (2000) and Panzarasa et al. (2001). In this respect Castelfranchi highlights the micro-macro link and the relationship of social functions and cognitive agents' mental representations mainly for two reasons:

First, and in particular, a theory of social function seems to be impossible to formulate without the sound knowledge of the relation between social functions and cognitive agents' mental representations, and second, social behaviour cannot be sufficiently explained without a theory of emergent social functions between cognitive (BDI) agents. Cognitive architectures are probably the most suitable way for a further analysis of this relationship. However, such an analysis also requires a sound consideration of emotion (Castelfranchi 2000: 6).

Investigation of the social functional components of emotion on three conceptual levels of analysis (micro-, meso, macro) can indeed be related to different approaches to (multi-)agent and artificial social systems quite intuitively. In these areas of inquiry, the concept of hybrid and multi-layered architectures has been brought a good step further (Castelfranchi 2000; Sloman/Logan 2000; Panzarasa/Jennings 2001/2002; Köhler/Rölke 2002). Layered architecture concepts generally build on a lower level of reactive, associative, and conditioned behaviour, on which further layers of increased deliberative capabilities and degrees of freedom rest. For example, Sloman's (2001) “CogAff” architecture is composed of reactive, deliberative, reflective, and self-conscious processes or layers, respectively.

Our approach to modelling the complex interdependencies and reciprocities of the social functions of emotion by taking a layered perspective is based on the multi-agent architecture MULAN that supplies a conceptually highly flexible framework in this respect. However, this framework is restricted by the SONAR architecture in order to achieve a clear separation of technical and conceptual models. MULAN concepts constitute a technical implementation of the agent concept that at the same time can already incorporate almost all aspects of an application model. For an improved separation of the different models we use the benefits and advantages of sociological concepts within the SONAR architecture (v. Lüde et al. 2003; Köhler et al. 2005). In particular, for each and every social unit – i.e., actors, processes, and social structures – a single SONAR agent is deployed and made available. The units' inherent logics are then described by multi-agent systems which are directly subordinated to the units in question.

In section 4 we will outline in more detail how a multi-level analysis of the social functional components of emotion can be modelled with the SONAR / MULAN architecture. But beforehand, we will further examine the role of norms and emotions in the exertion of social control.

3 Emotion and Social Control

Following the brief outline of the social functions of emotion, this section illustrates our perspective on social norms and their interactions with emotion in view of action incentives and social control.

3.1 Action Incentives

In order to lighten up the interrelation of emotion and social norms in view of the problem of structuration in artificial and natural societies, we first have to obtain clear concepts of how social norms and individual action relate to one another. If we assume that social norms constrain action and behaviour by stigmatising some options for action as more adequate than others, then we also can assume that structuration emerges in such a manner, that certain actions under certain situational conditions are not being implemented at all, and other options for action are constantly preferred by actors in such a way that robust “structuring practices” emerge (Knorr-Cetina 1981).

In this respect, Castelfranchi presupposes that in any case social norms themselves must be somehow consciously represented in order to function as action regulators. However, the *effects* of the social norms that are intended by a norm legislating entity do not have to be explicitly represented. Thus, the functional macro-structural effects of a social norm are intended from the viewpoint of a norm legislator, but unintended from the viewpoint of a norm-complying agent. The constrained and regulated agent only adopts the *function* of a social norm: "Normative behaviour has to be intentional and conscious: it has to be based on knowledge of the norm (prescription), but this does not necessarily imply consciousness and intentionality relative to all the *functions of the norm*" (Castelfranchi 2000: 23; italics original). This view is problematic for two reasons:

First, it implicitly assumes an intentional and omnipotent norm-issuing authority that satisfies the necessary prerequisites to establish social norms in view of their overall functional societal effects. The assumption of such a global system authority seems at least doubtful and rather pointless in an MAS context. Instead, we assume that social norms emerge (unlike laws created by global system authorities, e.g. the judiciary or a dictator) according to social evolutionary principles which might be useful to the autopoiesis of a system (cf. Horne 2001; Bendor/Swistak 2001).

Second, the question whether an agent acts consciously or unconsciously in compliance with social norms is in principle insignificant, because from a sociological point of view – which is indeed also shared by Castelfranchi – it is of paramount interest what *function* is carried out by standardised (and observable) behaviour. However, the question concerning the degree of consciousness of standardised (i.e., norm-abiding) behaviour only then arises, if one investigates the mechanisms which lead to the fact that social norms despite all cognitive competence, despite intentionality, and despite a supposed free will similarly cause the same observable behaviour. An explanation for this "foundational theoretical problem of the social sciences – the possibility of unconscious, unplanned emergent forms of cooperation, organisation and intelligence among intentional, planning agents" (Castelfranchi 2000: 5) (which is also known as Adam Smith's "invisible hand") is vainly looked for solely in the area of *conscious* norm oriented behaviour. Instead, we are convinced that social norms can in fact guide human behaviour without ever becoming consciously represented, e.g. in the form of imitative or habitualised behaviour.

This does not mean, however, that norms as such are per definitionem non-propositional entities that cannot be represented consciously. In another article together with Rosaria Conte, Castelfranchi champions the idea that "*norm-abiding* behaviour need not be based on the *cognitive processing of norms* (it might be simply due to imitation)" (Conte/Castelfranchi 1995: 187; italics original). If one presupposes that cognitive processing in this respect is meant to be conscious processing, then norm-abiding behaviour can in fact occur without conscious access to the norm. This contradicts Castelfranchi's statement concerning the functions of norms, namely that "norms, to work as norms, cannot remain unconscious in the addressee: the agent should *understand them as prescriptions and use them as such*" (Castelfranchi 2000: 23; italics original).

Nevertheless, we do strongly advocate the view that, regardless of the intentional nature of social norms, they have action regulating effects by way of their attachment to emotions, especially the social emotions. However, in doing so, we emphasise that the binding of norms (or of normative behaviour) to (social) emotions is a process that largely operates on an unconscious level. Here, the norm as such remains a sub-symbolic category, an action-script whose execution is, among other factors, fostered by accompanying emotions. In fact, these sub-symbolic mental structures may under certain circumstances be explicated and communicated to others and thereby become *intentional* and *social* objects, but they by no means have to, in order to be socially functional (cf. von Scheve/Moldt (2004) for details).

To further depict the interactions between norms and emotions it seems promising to conceptualise social norms as "mental objects", and thus primarily take into account their mental, cognitive, and emotionally decisive, but not necessarily conscious components. If we follow up this approach, social norms on the one hand become instances of the macro level because of their social, temporal, and spatial distribution. On the other hand, social norms simultaneously are instances of the micro level, because they are defined as properties or configurations of propositional attitudes (i.e., beliefs, desires, intentions) and cognitive representations and have profound influences on decision-making and action selection (cf. also Engel (2002), Carley (1986/1989), and Heckathorn (1989)).

This position of norms as mental objects is analogously put forward by Conte and Castelfranchi (1995: 192) who define social norms as "hybrid configurations of beliefs and goals". According to them, social norms, being directives or instructions which are represented as *beliefs*, substantially determine future actions of an agent by generating new goals*:* "they represent a powerful mechanism for inducing new goals in people's minds in a cognitive way" (Conte/Castelfranchi 1995: 189). However, the decisive questions in this respect, „how and why does a normative belief come to interfere with *x*'s decisions? What is it that makes her [an actor] responsive to norms concerning her? What is it that makes a normative belief turn into a normative goal?” (ibid. 192) are not answered satisfactory by Conte and Catselfranchi.

However, a sound answer to this question is of paramount significance if one wants to find a solution for the “foundational theoretical problem” mentioned by Castelfranchi. It is our conviction that emotions – especially in view of Jon Elster’s (1996/1998) concept – are of outstanding importance in this respect. In the following analysis, Elster’s concept shall serve as an addition to Conte and Castelfranchi’s position, since Elster rather delivers a definition *of certain qualities of social norms* than of the concept of social norms itself. Correspondingly, social norms can be described as follows (Elster 1999: 145f; see also Staller/Petta 2001):

1. Social norms are non-outcome-oriented phenomena. They can have unconditional imperative character but also conditional if they refer to past actions.
2. Social norms are shared with other members of a society or a social unit in which the process of sharing itself is also socially shared.
3. The third results from the second quality, namely that behaviour in compliance with a norm is subject to enforcement by other members of a social unit, also by means of sanctions (in order to achieve the definitional social sharing).

The following section shows how far social norms influence agents’ actions according to this position and which further-reaching emotion-related determinants of social action exist.

3.2 Control of Social Action

For the approach proposed here it is critical to examine the type of sanctioning in case of non-compliance to social norms. Particularly in economic theory, sanctions resulting from non-compliance are described as a withdrawal of material resources (Becker 1976; Axelrod 1986; Coleman 1990; Elster 1989). Material resources, however, are by no means the definitive or most influential objects of sanctions. Even more decisive in this respect is the fact that deviant agents interpret material sanctions also as a vehicle for the expression of negative emotions such as contempt, disdain, detestation, or disgust, and in consequence feel shame and/or guilt.

Shame in most cases will be interpreted as even worse because – in contrast to guilt – the perspective of the sanctioning agent is much more incorporated and accounted for. Furthermore, shame indicates a threat to an agent’s social bonds (Scheff 2003). Elster in this context explains that the *material* aspect of sanctions lies solely within the question of how much it costs the *punisher* to impose the sanction, and not on the question of how severe the sanctions are for the offender (Elster 1999: 146). To clarify: The higher the costs a punisher accepts to implement the intended sanctions, the more insistently aware is the offender of the negative emotions lying within these sanctions, and the more strongly the offender will feel the consequent shame. The amount of the punisher’s costs for sanctioning therefore signals to the offender the severity of the deviant behaviour. In many cases, punishers accept enormous costs that outreach by far the “damage” an offender has caused. But this surplus is by no means futile, since it is a way of making obvious the negative emotional meaning that comes along with the sanction and emphasises that the offender is expected to feel guilt or shame.

Frijda (1986) also takes a similar view at the interaction of social norms and emotion. He describes social rejection that results from emotional sanctions by means of shame or contempt as „severe punishment, […] most likely not merely because of its more remote adverse consequences“ (Frijda 1986: 351; Elster 1999: 147). Now, what consequences do these deterrents have for agents’ options to act?

Striving for emotional gratification, i.e. the motivation to seek encounters and interactions resulting in positive emotions and to avoid those resulting in negative emotions is considered a basic motivation of human behaviour. For example, Turner (1994) assumes that anxiety is one of the six primary motivational systems, whereas he defines anxiety as the “need to avoid a sense of disequilibrium with the environment“ (Turner 1994: 21). Emotions – in particular anxiety – can well relate to future actions by substantially affecting their actual planning. Giddens, for example, considers concerns over a loss of ontological security to be one central aspect in his theoretical framework: it is primarily the fear of the loss of ontological security and of facticity which serves as the central motivation of action (Giddens 1991). Other authors, e.g. Collins (1984) or Hammond (1991), who think of emotional gratification as a motivator of action that is directly scalable toward social aggregational contexts, assume that actors have an inborn need for positive emotional exchange processes, which may solidify to “interaction ritual chains” and contribute to the emergence of social structures (Collins 1981; Collins 2004).

Due to the interactions of emotion and social norms explained in the previous section, we can now further assume that in particular the emotions of shame and contempt serve as vehicles for the maintenance of social norms by generating *normative goals* (“n-goals”, as suggested by Conte and Castelfranchi (1995)) on the one hand and goals of avoidance of adverse consequences on the other hand. The goal of compliance with social norms therefore is *not* necessarily generated as a consequence of the anticipation of a loss of material resources through sanctions, but instead as a result of the fear of emotion-driven sanctions (by means of negative emotions such as, e.g., contempt, disdain, detestation, or disgust) that again result in negative social emotions, e.g., shame, guilt, or embarrassment in the offender.

However, the significance of emotion for the structural dynamics of social systems should not solely be clarified by pinpointing to the interactions with social norms. The significance of emotion for the control of social action and also for decision-making processes is just as decisive, without normative goals being necessarily generated. This point of view is also of importance for further-reaching examinations of the role of emotion in the emergence of social norms, which, however, cannot be done in this contribution (see von Scheve/Moldt 2004 for details).

4 Modelling of Emotional Agent Systems

We start with the presentation of some aspects of emotion that have been considered significant for computer science. We basically identify three different tendencies in modelling emotions within computer systems. Subsequently, we outline the role of norms and emotion in our modelling framework SONAR, whereas the model itself is being significantly developed by specific observations of emotion as a modelling subject.

4.1 Emotional Agents

The idea to improve artificial intelligence systems by taking into account emotions or functionally equivalent mechanisms is not new; it has its origins in the contributions of authors like Simon (1967), Sloman and Croucher (1981) or Minsky (1986). In the late 1980s first reviews of existing AI-models of emotions appeared (Dyer 1987; Pfeifer 1988). Until now, research on emotion within computer science has revealed three basic motivations to equip agents with artificial emotions: performance, human computer interaction, and simulation (Wehrle 1998; Picard 1997; Scheutz 2002).

So far, research on emotional agents has largely been concerned with either isolated entities or dyadic interaction settings (agent-agent / agent-user). In view of emotional agents being applied to distributed or multi-agent systems, we suppose that the foundational functional components of emotion (artificial and/or hybrid) in social aggregates, i.e., societies, teams, groups and organisations, have to be taken into account.

First efforts in this area have already treated the role and the potential of emotions in multi-agent systems, regarding problems like structuration, coordination, cooperation, and social control (Elliot 1992; Aubé/Senteni 1996; Gmytrasiewicz/Lisetti 2000; Staller/Petta 2001). Sociological research on emotion, which primarily investigates just these very problems, could contribute to considerably extend and optimise these approaches.

Almost all computer science models and systems that include emotions share the characteristic that they are based on psychological (and neuroscientific) theories. However, since emotions bear fundamental social components and significantly influence the social phenomena which are especially interesting to distributed artificial intelligence, a sociological consideration of emotions can open new and promising perspectives for computer science and at the same time also for the social sciences (cf. Sawyer 2003; Müller et al. 1998). It seems to be most debatable to either ignore social components of emotions in AI-systems or to insufficiently consider emotional effects on social phenomena and vice versa, which get especially relevant in distributed systems.

As far as computer science claims to consider those preconditions and consequences of emotions that are constitutive for the natural phenomenon and potentially serve the purposes of computer science, it cannot possibly miss to consider also the social functions of emotion.

4.2 Emotion and Norms in SONAR

Köhler and colleagues (2003) provide a modelling framework that allows conceiving social entities, i.e. social actors, social processes, and social structures, as “first-order objects” that can be modelled “side by side” simultaneously. The framework allows complete representations of direct interdependencies which are situated on the same layer of observation / abstraction. Internal properties of a particular social entity may stay entirely encapsulated from direct access by other social entities of the same layer.

Internal logics of each entity may differ from one another significantly – they are autonomous for each social entity. This way, any actor can, e.g., have an arbitrarily complex image of its (social) environment: this image is in turn depicted through a set of networked social entities, in this case represented by means of a multi-agent system based on SONAR. The system can be very simple for primitive agents, however, it can become highly complex for agents exhibiting higher degrees of social differentiation.

For example, imagine looking at the mind of some sociologist and his internal representation of the external and internal world in its entire complexity – including all contents and any probable inconsistency. This would be a theoretical example without claiming to model a real person; rather, the designer can arbitrarily simplify or extend the model according to the requirements of the task. The same is also true for the processes that are established between social actors and agents, as well as for fixed, but in the long run alterable social structures. Therefore, social entities in this framework *in themselves* contain the necessary references to other entities, i.e. these references do not have to be modelled separately.

To design a specific system, the designer is advised to deliberately choose specific aspects of a system which are supposed to be central and most relevant for the modelling goals. The SONAR models then specify exactly the most important elements of the chosen model. On the other hand, MULAN so far provides so far the technical framework for implementing the principal concepts of agents and multi-agent systems like autonomy, mobility, cooperation and adaptivity. The SONAR architecture enables modelling of the internal representation of actors in terms of a multi-agent system. Furthermore, micro-, meso- and macro-layers are being modelled, whereas micro is to be understood as an actor, meso as an interaction or process, and macro as a social structure. Figure 1 illustrates the interplay of two autonomous social entities, in this case an actor and the social process an actor is involved in. The incorporation into a social structure takes place by means of the process which is involved in the stabilisation and reproduction (or development) of new structures. Conceptually, we have to deal with the same sort of connection between both parts of the model (synchronous channels).

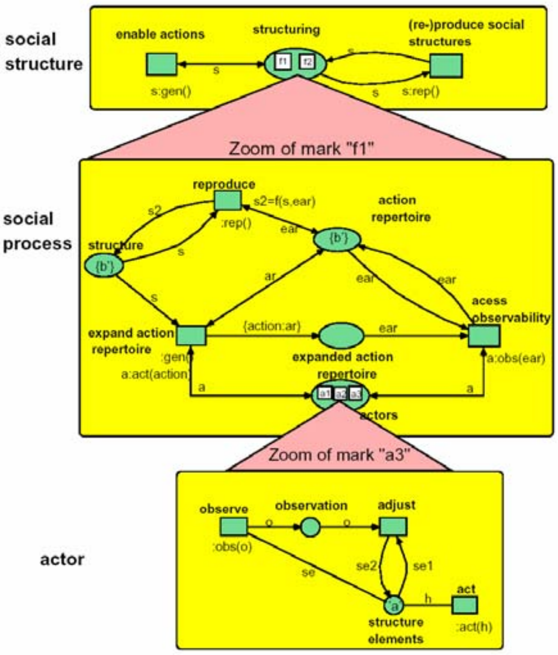


Fig. 1: A reference net representing actor, social process, and social structure (cf. v. Lüde et al. 2003)

In figure 1 a powerful variant of Petri nets – a reference net (Kummer 2002) – is shown. Rectangles (transitions) represent activities or actions, while circles (places) denote resources that can be available or not, or conditions that may be fulfilled. Arcs determine the specific context of the transition. Thus, arcs that are directed from transitions to places can be interpreted as preconditions for actions, whereas arcs that are directed from places to transitions represent actions' outcomes. A transition that fires (or: an action that is carried out) removes resources or conditions (for short: tokens) from places and inserts them into other places. A peculiar property of reference nets is the possibility that tokens on the places of a net can also be reference nets again (or either arbitrary Java-objects).

Surrounding nets are called *system nets*, those on the places *object nets* (Valk 1998). Object net and system net are synchronised by means of synchronous channels, whereby one of the nets is expected to *reference* the other one. For example, actions of an actor are synchronised with a social process by means of two transitions that constitute the synchronous channel: *observe* and *act* in actor net, here modelled as object nets, and *access visibility* and *expand action repertoire* in the social process net, in this case the system net for the actor. Tokens a1, a2 and a3 can be refined into actor nets.

It should be pointed out that this kind of refinement succeeds by means of zooming that is performed on references to the relevant actor nets. Any social entity is for itself essentially autonomous. The mutual relationship can be principally seen as symmetrical, even though modelling can imply a (rational) hierarchical structure. The concepts of respective observation and action enter the interaction via the social process entity. The adjustment succeeds locally, without direct interaction with the environment. The inscriptions of social process signify the embedding into the social structure, though these more special aspects had to be omitted here. Details can be found in either v. Lüde et al. (2003) or in Köhler et al. (2005).

Although sociological terms, such as acknowledgement, observation, action, actor, etc., are widely in use within SONAR models, the role of emotions has hardly been explicitly taken into account so far. This proves to be obstructive for modelling emotions and norms in the context of SONAR. Hence we propose some simple but fundamental enhancements to SONAR.

We can still conventionally represent the standard approaches; however we are going to treat all aspects concerning emotion separately. This approach demands an explicit decision to be made by the model designer in order to precisely determine what can and has to be classified as belonging to an emotion proper. Hence, we build a separate emotion model besides the conventional view which is rather rational and utility-based.

The problem of connecting the separate models will be solved in a simple and homogenous manner, based on the applied modelling techniques: existing models are to be completed with further emotion models. The statements of the sociological (and other) emotion theories either deal with pure emotion-based interdependences, in which case they are to be integrated only in the emotion models, or emotions are additionally claimed to influence social structures and processes. In the latter case, linkage to the rational elements of the model is accomplished by means of synchronous channels, thus creating some kind of corresponding “parallel world”. Modelling all social entities as individual nets allows arbitrarily setting them in relationship to each other through references (in the sense of a “pointer”). Through synchronous channels any possible linkage of actions from different parts of the model can be represented.

Conceptually, we therefore reach a definite separation of emotion from the so far non-emotional elements of the sociological models. Thus, emotion can be explicitly split down into its relevant integral components. Obviously, the general system complexity cannot be overcome by these means although it is distributed to different levels. Their integration demands even further efforts, since the linkages between the different levels have to be created explicitly (Here we can, e.g., proceed by applying solutions of analogous problems which can be found, for example, in the combination of different viewpoints in models created with the unified modelling language (UML).) Although at some point in the modelling process integration is necessary and essential, we currently favour the advantages resulting from an explicit separation of the models, in particular the intuitive simplicity that comes along with these detached points of view. These advantages can be clarified by referring to UML models of highly complex systems.

Smaller and less complex models can often be implemented faster with those programming languages that do not distinguish between different points of view. Though, the separation of different points of view in order to handle system complexity is a common method used in computer science. Apart from complexity issues in computer science, also sociological models are highly complex thought systems that may demand even more requirements in view of flexibility as this is generally the case for construction-oriented computer science models. Therefore, a separation into different layers seems to be essential not only for sociological emotion theories but also in view of a sharper examination of different analytical layers.

To summarise: While SONAR incorporates a division into actor, social process, and social structure, and at the same time facilitates corresponding points of view (while also operating with special model patterns, e.g. *actor* is supposed to involve observation, adoption and action (see figure 1)), we have proposed a further dimension which is especially significant because of its content-based and therefore application-related qualities: the supplementary modelling of emotion.

The possibility to design discrete emotions and their specific components as explicit states or processes which are integrated into the existing (rational) models still remains unelaborated, but is a goal for future work. Although the approach presented here allows a sharp distinction, it does not enforce it, since profound experiences with applications are still missing and restrictions of the flexibility might be possible. However, these restrictions should only arise from the requirements of the specific theories of emotion, which can provide the appropriate and substantial arguments.

5 Discussion

Finally we briefly evaluate the modelling approach presented above as well as the possibilities it bears in view of modelling the interrelation of emotion and norms:

Generally, by considering emotions the capabilities of simulating and modelling sociological theory are crucially improved. A differentiated presentation facilitates the reduction of complexity concerning the adopted view and the particular system, supported by the capacities of UML. The technical implementation of emotional mechanisms within the SONAR / MULAN-architecture is still under current development.

This simulation environment can be subsequently used to improve multi-agent systems, for example in view of alternative coordination solutions. However, evidence must yet be presented, whether modelling emotions can ever enable such solutions. Still, following Minsky (1986), we take it for granted, that intelligent systems need a replication of the (social) functions of emotion.

The interaction between emotion and deliberation/cognition can now be investigated on the basis of explicitly separated components of a model, whereby specific sociological questions can also be addressed. The acknowledged separation primarily aims at analytical clarity and the explicit modelling of interactions and interdependencies, and it does not target the explanation of natural phenomena.

The explicit representation of emotion on the basis of analytical models that are still to be developed can in addition be used and further scrutinised in the context of human-computer interaction.

The usage of SONAR with its conceptual distinction of actors, processes, and structures as well as patterns, that are to be specified still more precisely, makes it possible to reflect the central social theoretical concepts, such as the interrelation of norms and emotion, especially in the contexts mentioned above.

There exists a further architecture that is directly embedded into a FIPA-conform agent-system framework (CAPA) (Duvigneau et al. 2003). Its conceptual foundation MULAN provides software-technical support of application-related concepts, thus sparing the transfer efforts of the model designer.

The dependences between norms and emotions can now be examined on different levels. Thus, on the structural layer we can combine norms either with the modelling concepts, which largely disregard emotions, or with those that consider emotions on all layers of a model (actor, process, and structure).

This flexibility permits to consider different theories of emotion *simultaneously* and to different extents. At this point, the existing emotional agent architectures, which are primarily concerned with modelling actor- and process-layer (e.g., TABASCO (Staller/Petta 1998)), can be picked up, integrated, and extended with the sociological aspects presented above.

The decomposition into actor, process, and structure also supports the constitution of the examined emotion theories that expand from the neurological and cognitive to the sociological points of view. Moreover, this separation may crucially simplify the formulation of an interdisciplinary theory of emotion, since it acknowledges the interconnections between sociological, psychological, and probably also neurological theories (see von Scheve/Moldt 2004).

Acknowledgements

The first author acknowledges a grant from the Research Group “Emotions as Bio-Cultural Processes“, Center for Interdisciplinary Research (ZiF), University of Bielefeld, Germany.

References

J.A. Abe and C.E: Izard. The Developmental Functions of Emotions: An Analysis in Terms of Differential Emotions Theory. *Cognition & Emotion*, 13(5): 523-549, 1999.

J.R. Averill. The Structural Bases of Emotional Behavior. In: M.S. Clark (Ed.). *Emotion*. Review of Personality and Social Psychology, Vol. 13. Newbury Park/CA: Sage, 1-24, 1992.

M. Aubé and A. Senteni. Emotions as Commitments Operators: A Foundation for Control Structure in Multi-Agents Systems. In: W. van de Velde and J.W. Perram (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, 1996.

R. Axelrod. An evolutionary approach to norms. *American Political Science Review*, 80: 1095-1111, 1986

A.L. Bazzan, D.F. Adamatti and R. Bordini. Extending the Computational Study of Social Norms with a Systematic Model of Emotions. In: G. Bittencourt and G.L. Ramalho (Eds.). *Advances in Artificial Intelligence*. Proceedings of the 16th Brazilian Symposium on Artificial Intelligence (SBIA’02), Porto de Galinhas/Recife, Brazil, November 11-14. LNCS Vol. 2507. Berlin: Springer, 108-117, 2002.

J. Bendor and P. Swistak. The Evolution of Norms. *American Journal of Sociology*, 106(6): 1493-1545, 2001

G. Becker. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press, 1976.

J.T. Cacioppo, G.G. Berntson, J.T. Larsen, K.M. Poehlmann and T.A. Ito. The Psychophysiology of Emotion. In: R. Lewis and J.M. Haviland-Jones (Eds.). *Handbook of Emotions*, 2nd Ed. New York: Guilford Press, 173-191, 2000.

D. Cañamero (Ed.). *Emotional and Intelligent: The Tangled Knot of Cognition*. Proceedings of the 1998 AAAI Fall Symposium. Technical Report FS-98-03. Menlo Park/CA: The AAAI Press, 1998.

K. Carley. An Approach for Relating Social Structure to Cognitive Structure. *Journal of Mathematical Sociology*, 12(2): 137-189, 1986.

K. Carley. The Value of Cognitive Foundations for Dynamic Social Theory. *Journal of Mathematical Sociology*, 14(2/3): 171-208, 1989.

C. Castelfranchi. The Theory of Social Functions: Challenges for Computational Social Science and Multi-Agent Learning. *Journal of Cognitive Systems Research*, 2(1): 5-38, 2000.

G.L. Clore, N. Schwarz and M. Conway. Affective Causes and Consequences of Social Information Processing. In: R.S. Wyer and T.K. Srull (Eds.). *Handbook of Social Cognition*. 2nd Ed. Hillsdale/NJ: Lawrence Erlbaum, 323-417, 1994.

G.L. Clore. Why Emotions Are Felt. In: P. Ekman and R.J. Davidson (Eds.). *The Nature of Emotion*. New York: Oxford University Press, 103-111, 1994.

J.S. Coleman. *Foundations of Social Theory*. Cambridge/MA: Harvard University Press, 1990.

R. Collins. On the Microfoundations of Macrosociology. *American Journal of Sociology*, 86(5): 984-1014, 1981.

R. Collins. The Role of Emotion in Social Structure. In: K.R. Scherer and P. Ekman (Eds.). *Approaches to Emotion*. Hillsdale/NJ: Lawrence Erlbaum, 385-396, 1984.

R. Collins. *Interaction Ritual Chains*. Princeton/NJ: Princeton University Press, 2004.

R. Conte and C. Castelfranchi. Norms as Mental Objects – From Normative Beliefs to Normative Goals. In: C. Castelfranchi and J.-P. Müller (Eds.). *From Reaction to Cognition*. Selected Papers from the 5th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW ’93), 25.-27. August, Neuchatel/CH. LNAI Vol. 957. Berlin: Springer, 186-196, 1995.

L. Cosmides and J. Tooby. Evolutionary Psychology and the Emotions. In: R. Lewis and J.M. Haviland-Jones (Eds.). *Handbook of Emotions*, 2nd Ed. New York: Guilford Press, 91-115, 2000.

F. Dignum, D. Morley, E.A. Sonenberg and L. Cavedon. Towards socially sophisticated BDI agents. In: *Proceedings of the 4th International Conference on Multiagent Systems (ICMAS’00)*. Boston/MA: IEEE, 2000.

U. Dimberg. Psychophysiological Reactions to Facial Expressions. In: U. Segerstrale and P. Molnar (Eds.). *Nonverbal Communication: Where Nature Meets Culture*. Mahwah/NJ: Lawrence Erlbaum, 47-60, 1997.

E. Durkheim. *Die elementaren Formen des religiösen Lebens*. Frankfurt am Main: Suhrkamp, (1994)[1912].

M. Duvigneau, D. Moldt and H. Rölke. Concurrent Architecture for a Multi-agent Platform. In: F. Giunchiglia, J. Odell and G. Weiß (Eds.). *Agent-Oriented Software Engineering III*. Proceedings of AOSE’02. LNCS Vol. 2585. Berlin: Springer, 59-72, 2003.

M.G. Dyer. Emotions and their computations: Three computer models. *Cognition & Emotion*, 1(3): 323-347, 1987.

C.D. Elliot. Using the Affective Reasoner to Support Social Simulations. *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI’93)*, Vol. 1. Los Altos/CA: Morgan Kaufmann, 194-200, 1993.

C.D. Elliot. *The Affective Reasoner. A Process Model of Emotions in a Multi-Agent System*. PhD Thesis, Technical Report 32, Intitute for the Learning Sciences, Northwestern University, 1992.

J. Elster. Emotions and Economic Theory. *Journal of Economic Literature*, 36(1): 47-74, 1989.

J. Elster. Rationality and the Emotions. *The Economic Journal*, 106(438): 1386–1397, 1996.

J. Elster. *Alchemies of the Mind. Rationality and the Emotions*. Cambridge: Cambridge University Press, 1999.

P. Engel. Intentionality, Normativity and Community. *Facta Philosophica* 4(1): 25-49, 2002.

J. Fix. *Emotionale Agenten*. Diploma Thesis. University of Hamburg, Computer Science Department, 2004.

N.H. Frijda. *The Emotions*. Cambridge: Cambridge University Press, 1986.

J. Gerhards. *Soziologie der Emotionen: Fragestellungen, Systematik u. Perspektiven*. Weinheim: Juventa, 1988.

A. Giddens. *Modernity and Self-Identity*. Cambridge: Polity Press, 1991.

P.J. Gmytrasiewicz and C.L. Lisetti. Using Decision Theory to Formalize Emotions for Multi-Agent Systems. *Proceedings of the 2nd ICMAS’00 Workshop on Game Theoretic and Decision Theoretic Agents*. Boston/MA, 2000.

G. Hatano, N. Okada and H. Tanabe (Eds.). Affective Minds. Proceedings of the 13th Toyota Conference. Amsterdam: Elsevier, 2000.

D.D. Heckathorn. Cognitive Science, Sociology, and the Theoretic Analysis of Complex Systems. *Journal of Mathematical Sociology*, 14(2/3): 97-110, 1989.

A.R. Hochschild. Emotion Work, Feeling Rules, and Social Structure. *American Journal of Sociology*, 85(3): 551-575, 1979.

A.R. Hochschild. *The Managed Heart*. Berkeley/CA: University of California Press, 1983.

M. Holodynski and W. Friedlmeier. *Development of Emotions and Their Regulation*. Boston/MA: Kluwer Academic, 2005[in press].

C. Horne. Sociological Perspectives on the Emergence of Norms. In: M. Hechter and K.-D. Opp (Eds.). *Social Norms*. New York: Russell Sage Foundation, 3-34, 2001.

Keltner, D.; Gross, J.J. (1999): Functional Accounts of Emotion. *Cognition & Emotion*, 13(5): 467-480.

D. Keltner and J. Haidt. Social Functions of Emotion at Four Levels of Analysis. *Cognition & Emotion*, 13(5): 505-521, 1999.

T.D. Kemper. *A Social Interactional Theory of Emotions*. New York: Wiley & Sons, 1978.

K.D. Knorr-Cetina. Introduction: The micro-sociological challenge of macro-sociology. In: K.D. Knorr-Cetina and A.V. Cicourel (Eds.). *Advances in social theory and methodology*. Boston/MA: Routledge & Kegan Paul, 1-47, 1981.

M. Köhler, D. Moldt and H. Rölke. Modeling the Structure and Behaviour of Petri Net Agents. In: J.M. Colom and M. Koutny (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, 2001.

M. Köhler, D. Moldt and H. Rölke. A Discussion of Social Norms with Respect to the Micro-Macro Link. In: *Proceedings of the 2nd International Workshop on Regulated Agent-Based Social Systems: Theories and Applications* (RASTA'03), June 24.-26., Edinburgh, 2003.

M. Köhler, D. Moldt, H. Rölke and R. Valk. Linking micro and macro descriptions of scalable social systems using reference nets. In: K. Fischer and M. Florian (Eds.). *Socionics: Its Contribution to the Scalability of Complex Social Systems.* LNCS. Heidelberg: Springer, 2005[in press].

M. Köhler and H. Rölke. Modelling the Micro-Macro Link: Towards a Sociologically Grounded Design of Multi Agent Systems. In: C. Jonker, G. Lindemann and P. Panzarasa (Eds.). *Proceedings of the Workshop Modelling Artificial Societies and Hybrid Organizations (MASHO'02),* 2002.

O. Kummer. *Referenznetze*. Berlin: Logos-Verlag, 2002.

E.J. Lawler, S.R. Thye and J. Yoon. Emotion and Group Cohesion in Productive Exchange. *American Journal of Sociology*, 106(3): 616-657, 2000.

R. von Lüde, D. Moldt and R. Valk (Eds.). *Sozionik. Modellierung soziologischer Theorie*. Münster: Lit-Verlag, 2003.

D. Moldt and C. von Scheve. Emotions in Hybrid Social Aggregates. In: M. Herczeg, W. Prinz and H. Oberquelle (Eds.). *Mensch & Computer 2002. Vom interaktiven Werkzeug zu kooperativen Arbeits- und Lernwelten*. Stuttgart: Teubner, 343-352, 2002.

M. Minsky. *The Society of Mind*. New York: Simon & Schuster, 1986.

H.J. Müller, T. Malsch and I. Schulz-Schaeffer. Socionics: Introduction and Potential. *Journal of Artificial Societies and Social Simulation*, 1(3), 1998. http://www.soc.surrey.ac.uk/JASSS/1/3/5.html

K. Oatley and J.M. Jenkins. *Understanding Emotions*. Oxford: Basil Blackwell, 1996.

P. Panzarasa and N.R. Jennings. The Organisation of Sociality: A Manifesto for a New Science of Multi-Agent Systems. *Proceedings of the 10th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'01),* Annecy, France, 2001.

P. Panzarasa and N.R. Jennings. Social Influence, Negotiation and Cognition. *Simulation Modelling Practice and Theory,* 10(5-7): 417-453, 2002.

P. Panzarasa, T.J. Norman and N.R. Jennings. Social mental shaping: Modelling the impact of sociality on autonomous agents' mental states. *Computational Intelligence*, 17(4): 738-782, 2001.

R. Pfeifer. Artificial Intelligence Models of Emotion. In: V. Hamilton, G.H. Bower and N.H. Frijda (Eds.). *Cognitive Perspectives on Emotion and Motivation*. Dodrecht: Kluwer Academic, 287-320, 1998.

R.W. Picard. *Affective Computing*. Cambridge/MA: The MIT Press, 1997.

N.J. Saam and A. Harrer. Simulating Norms, Social Inequality, and Functional Change in Artificial Societies. *Journal of Artificial Societies and Social Simulation*, 2(1), 1999. http://www.soc.surrey.ac.uk/JASSS/2/1/2.html

R.K. Sawyer. Artificial Societies. Multiagent Systems and the Micro-Macro Link in Sociological Theory. *Sociological Methods & Research*, 31(3): 325-363, 2003.

R.K. Sawyer. Social Explanation and Computational Simulation. *Philosophical Explorations*, 7(3): 219-231, 2004.

T.J. Scheff. Shame in Self and Society. *Symbolic Interaction*, 26(2): 239-262, 2003.

M. Scheutz. Agents with or without Emotions? In: S. Haller and G. Simmons (Eds.). *Proceedings of the 15th International Florida Artificial Intelligence Symposium* (FLAIRS'02). Menlo Park/CA: The AAAI Press, 89-94, 2002.

C. von Scheve and D. Moldt. Emotion: Theoretical Investigations and Implications for Artificial Social Aggregates. In: G. Lindeman, D. Moldt and P. Paolucci (Eds.): *Regulated Agent-Based Social Systems*. LNAI Vol. 2934. Berlin: Springer, 189-209, 2004.

N. Schwarz. Feelings as information: Informational and motivational functions of affective states. In: R.M. Sorrentino and E.T. Higgins (Eds.). *Handbook of Motivation and Cognition*. Foundations of Social Behavior, Vol. 2. New York: Guilford Press, 527-561, 1990.

H.A. Simon. Motivational and emotional controls of cognition. *Psychological Review*, 74: 29-39, 1967.

A. Sloman. Varieties of Affect and the CogAff Architecture Schema. In: *Agents and Cognition. Proceedings of the AISB'01 Symposium on Emotion, Cognition, and Affective Computing*. York: SSAISB, 39-48, 2001.

A. Sloman and M. Croucher. Why robots will have emotions. *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI'81)*. Vancouver, British Columbia, 197-202, 1981.

A. Sloman and B. Logan. Evolvable Architectures for Human-Like Minds. In: G. Hatano, N. Okada and H. Tanabe (Eds.). *Affective Minds*. Proceedings of the 13th Toyota Conference, Nov.-Dec. 1999, Nagoya/Japan. Amsterdam: Elsevier, 169-182, 2000.

C.A. Smith and L.K. Pope. Appraisal and Emotion. The Interactional Contributions of Dispositional and Situational Factors. In: M.S. Clark (Ed.). *Emotion and Social Behavior*. Review of Personality and Social Psychology, Vol. 14. Newbury Park/CA: Sage, 32-62, 1992.

A. Staller and P. Petta. Towards a tractable appraisal-based architecture for situated cognizers. In: D. Cañamero, C. Numaoka and P. Petta (Eds.). *Grounding Emotions in Adaptive Systems*. Workshop Notes of the 5th International Conference of the Society for Adaptive Behaviour (SAB'98). Zurich, Switzerland, 56-61, 1998.

A. Staller and P. Petta. Introducing Emotions into the Computational Study of Social Norms: A First Evaluation. *Journal of Artificial Societies and Social Simulation*, 4(1), 2001. http://soc.surrey.ac.uk/JASSS/4/1/2.html

R. Trappl, P. Petta and S. Payr (Eds.). *Emotions in Humans and Artefacts*. Cambridge/MA: The MIT Press, 2003.

J.H. Turner. A General Theory of Motivation and Emotion in Human Interaction. *Österreichische Zeitschrift für Soziologie*, 19(1): 20-35, 1994.

J.H. Turner. *On the Origins of Human Emotions. A Sociological Inquiry into the Evolution of Human Affect*. Stanford: Stanford University Press, 2000.

J.H. Turner. *Face to Face. Toward a Sociological Theory of Interpersonal Behavior*. Stanford/CA: Stanford University Press, 2002.

R. Valk. Petri Nets as Token Objects. An Introduction to Elementary Object Nets. In: J. Desel and M. Silva (Eds.). *Proceedings of Application and Theory of Petri Nets*. Lisbon, Portugal. LNCS Vol. 1420. Berlin: Springer, 1-25, 1998.

T. Wehrle. Motivations Behind Modeling Emotional Agents: Whose emotions does your robot have? In: D. Cañamero, C. Numaoka and P. Petta (Eds.). *Grounding Emotions in Adaptive Systems*. Workshop Notes of the 5th International Conference of the Society for Adaptive Behaviour (SAB'98). Zurich, Switzerland, 1998.

Normative KGP Agents: A Preliminary Report

Fariba Sadri and Francesca Toni[⋆]
[⋆]Department of Computing,
Imperial College London, UK.
{fs,ft}@doc.ic.ac.uk

Kostas Stathis[†]
[†]Department of Computing,
City University London, UK.
kostas@soi.city.ac.uk

Abstract

We extend the logical model of agency known as the KGP model, to support agents with normative concepts, using obligations and prohibitions as examples. The proposed framework illustrates how to integrate normative concepts and roles within the KGP model in such a way that these concepts can evolve dynamically. Furthermore, we illustrate how these concepts can be combined with the existing capabilities of KGP agents in order to plan and react to changes in the environment. Our approach gives an executable specification of normative concepts that can be used directly for prototyping applications.

1 Introduction

Programmers that develop complex distributed systems based on autonomous agents often need to find ways of decentralising the functionality of the whole system by distributing responsibility to the parts, while still ensuring that interactions of the whole are coherent and coordinated. Whether we deal with autonomous robots that plan, a traffic or a file-sharing system, or any other similar application, it is becoming increasingly recognised that the resolution of the underlying problems lies with developing frameworks that are based on the notion of social agency (Huhns and Singh, 1998).

The basic idea behind the notion of social agency is to use abstractions from such diverse fields as sociology, computing science, organisational theory, and law in order to specify (Jones and Sergot, 1993) and then implement complex organisations of agents referred to as *artificial societies* (Padget, 2001). Such an implementation seeks to apply social agents in practical applications where the formation of open societies (Artikis and Pitt, 2001) are envisaged to be regulated by *norms*. The notion of a norm is important here in that member agents of a society (Toni and Stathis, 2002) must have the capability to reason with norms and they must be capable of communicating norms to other member agents. The problem then reduces to resolving the issue of how to develop normative autonomous agents (Dignum, 1999) for supporting practical applications based on artificial societies.

To develop normative agents programmers often rush into implementing the rules governing an artificial society without properly understanding the subtleties that underlie their specification. Motivated by this observation a number of researchers, e.g. see (Castelfranchi et al., 1999; Jones and Sergot, 1996; Carabelea et al., 2004), are seeking to understand the modalities required to specify the norms of an artificial society as a separate issue from their implementation. Deontic concepts such as obligation, prohibition, permission, rights, power, and entitlement, are currently being scrutinised and analysed in detailed formal frameworks that have been produced as a result. Much of the work in this effort, however, can only be used as a guide to implementations, as it abstracts away from the computational characterisation of the resulting specifications and, more importantly, from the way these normative concepts are to be used during the operation of the agent in an artificial social environment.

Motivated by this last observation, this work seeks to complement the specification effort previously described by presenting a framework that illustrates how normative concepts, such as obligations and prohibitions, can be used by an agent while it reasons, reacts, plans, and communicates in the context of an artificial society. We develop this framework by building upon existing work with the KGP (Knowledge, Goals, Plans) model of agency (Kakas et al., 2004b) which we have successfully implemented in the prototype agent platform PROSOCS (Stathis et al., 2004). The contribution of the work is to show how to extend the KGP model with normative concepts, thus providing a framework where we can develop agents who reason about norms that govern, not only their own behaviour, but also the behaviour of other agents. One major advantage of

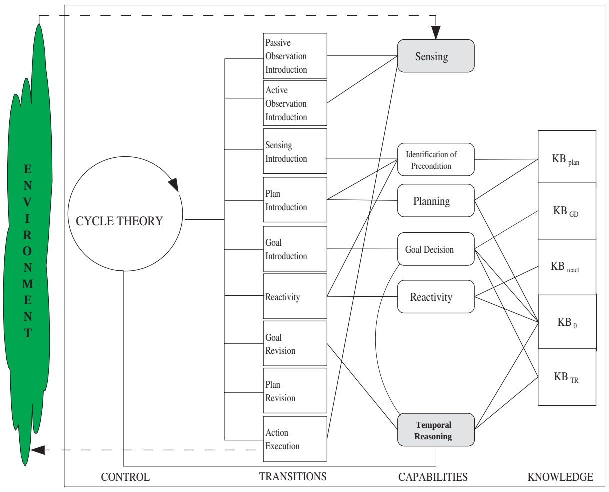


Figure 1: The KGP Model of Agency.

our proposed approach is that the extension to the KGP model is smooth, without requiring new capabilities, but simply extending the existing knowledge bases with which the agents reason when they operate. Another distinguishing feature of our approach that we consider as another major advantage is that the normative rules that we propose can be looked at in many different ways: they are executable specifications, they are directly implementable and, within the declarative and operational model of the KGP agents, they force agents to exhibit the expected behaviour, conformant with the specification.

We structure the paper as follows. In the next section we summarise the main features of the KGP model. Then, in the following section we outline how to extend the current model so that we can accomodate normative concepts such as obligations. We then exemplify the extended model with a scenario, showing how an agent can use the specifications to produce norm-governed baheaviour. Related work is discussed in the penultimate section, while in the final section we summarise the main features of the work and outline our plans for future work.

2 The KGP Model

Here we briefly summarise the KGP model for agents, which is depicted in Figure 1. We focus on the components relevant to this paper, see (Kakas et al., 2004b; Bracciali et al., 2004) for any additional details.

The model relies upon:

- an internal (or mental) state,
- a set of reasoning capabilities, supporting planning, temporal reasoning, identification of preconditions of actions, reactivity and goal decision,
- a sensing capability,
- a set of transition rules, defining how the state of the agent changes, and defined in terms of the above capabilities,
- a set of selection functions, to provide appropriate inputs to the transitions,
- a cycle theory, for deciding which transitions should be applied when, and defined using the selection functions.

Internal state. This is a tuple

$$\langle KB, Goals, Plan, TCs \rangle,$$

where:

- **KB** describes what the agent knows of itself and the environment and consists of separate modules supporting the different reasoning capabilities, including
  - KB_{plan}, for Planning,
  - KB_{pre}, for the Identification of Preconditions of actions,
  - KB_{react}, for Reactivity, and
  - KB_0, for holding the (dynamic) knowledge of the agent about the external world in which it is situated (including past communications), and perceived through its sensing capability.

  Syntactically, KB_{plan} and KB_{react} are abductive logic programs with constraint predicates (see (Bracciali et al., 2004)), and KB_{pre} is a logic program.
- **G**oals is a set of properties that the agent wants to achieve, each one explicitly time-stamped by a time variable. Goals may also be equipped with a temporal constraint (given in TCs) binding the time variable and constraining when the goals are expected to hold. Goals may be *mental* or *sensing*. Both can be observed to hold (or not to hold) via the Sensing capability. In addition, mental goals can be brought about actively by the agent by its Planning capability and its actions.

- **P**lan is a set of actions scheduled in order to satisfy goals. Each is explicitly time-stamped by a time variable and possibly equipped with a temporal constraint (given in TCs), similarly to $Goals$, but constraining when the action should be executed. Actions are partially ordered, via their temporal constraints. Each action is also equipped with the preconditions for its successful execution, determined by the Identification of Preconditions capability. Actions may be *physical*, *communicative*, or *sensing*.

- TCs is a set of constraint atoms (referred to as *temporal constraints*) in some given underlying constraint language. We assume that the constraint predicates include $<, \leq, >, \leq, =, \neq$. These constraints specify when goals are to hold and when actions are to be executed, and they are extended and instantiated as the agent operates.

Reasoning capabilities. These include:

- Planning, which generates partial plans for sets of goals. It provides (temporally constrained) sub-goals and actions designed for achieving the input goals.

- Reactivity, which reacts to perceived changes in the environment, by replacing (some) goals in $Goals$ and actions in $Plan$ with (possibly temporally constrained) goals and actions.

- Identification of Preconditions for action execution.

Sensing capability. This links the agent to its environment, by allowing to observe that properties hold or do not hold, and that other agents have executed actions in the past. It also allows agents to receive communication from other agents.

Transitions. The state of an agent evolves by applying transition rules, which employ capabilities. The transitions include:

- *Passive Observation Introduction* (POI) changes KB_0 by introducing unsolicited information coming from the environment or communications received from other agents. It calls the Sensing capability.

- *Plan Introduction* (PI) changes part of the $Goals$ and $Plan$ and TCs of a state, according to the output of the Planning capability. This transition uses also the Identification of Preconditions capability, in order to equip each planned action A with the set of preconditions for the successful execution of A.

- *Reactivity* (RE) is responsible for updating the current state of the agent by adding the goals and actions, together with any additional temporal constraints, returned by the Reactivity capability. As with PI, this transition too uses the Identification of Preconditions capability, in order to equip each new action A with the set of preconditions for the successful execution of A.

- *Action Execution* (AE) is responsible for executing all types of actions, thus changing the KB_0 part of KB by adding evidence that actions have been executed. It calls the Sensing capability for the execution of sensing actions.

Cycle. The behaviour of an agent is given by the application of transitions in sequences, repeatedly changing the state of the agent. These sequences are not determined by fixed cycles of behaviour, as in conventional agent architectures, but rather by reasoning with cycle theories. These are logic programs with priorities (see (Kakas et al., 2004a)), defining preference policies over the order of application of transitions, which may depend on the environment and the internal state of a agent.

In the remainder of this section, we give some details on the state of agents in the original KGP model, to provide the necessary background to incorporate normative concepts.

Goals. A goal is a *timed fluent literal* $l[t]$, where l refers to a (positive or negative) property that the agent wants to hold and t is the *time* of the goal, namely a variable, implicitly existentially quantified within the overall state of the agent. An example is $has\_driving\_licence(ag_1, t_1)$, indicating that agent ag_1 wants to have a driving licence at time t_1. This time may be constrained by TCs in the state, e.g. TCs may contain $10 < t_1 < 20$.

Actions. An action is a *timed operator* $a[t]$, where a refers to the operator of the action, and t is the *execution time* of the action, namely a variable, implicitly existentially quantified within the overall state of the agent. An example is $pay\_fine(ag_1, t_2)$, indicating that agent ag_1 wants to perform an act of paying a fine at time t_2. Again, this time may be constrained by TCs in the state, e.g. TCs may contain $t_2 < t_1$.

KB_0**.** This records the actions which have been executed (by the agent or by others) and their time of execution as well as the properties (i.e. fluents and their negation) which have been observed, possibly concerning other agents, and the time of the observation. Formally, KB_0 contains assertions of the form:

- $executed(a[t], \tau)$ where $a[t]$ is a timed operator and τ is a time constant, meaning that action a has been executed at time $t = \tau$, by the agent holding the KB_0.
- $observed(ag', a[ag, \tau], \tau')$ where τ is a time constant, meaning that action a has been executed at time τ by the agent ag, and this has been observed by the agent ag' (different from ag) holding the KB_0.
- $observed(l[t], \tau)$ where $l[t]$ is a timed fluent literal and τ is a time constant, meaning that the property l has been observed to hold at time $t = \tau$, by the agent holding the KB_0.

KB_{plan} **and Planning capability.** KB_{plan}, KB_{react}, and KB_{pre} are all specified within the framework of the event calculus (EC) for reasoning about actions, events and changes (Kowalski and Sergot, 1986). Below, we give the abductive logic program KB_{plan} and the logic program KB_{pre}.

In a nutshell, the EC allows to write meta-logic programs which "talk" about object-level concepts of *fluents*, *events* (that we interpret as *action operations*), and *time points*. The main meta-predicates of the formalism are: $holds\_at(F, T)$ (a fluent F holds at a time T), $clipped(T_1, F, T_2)$ (a fluent F is clipped - from holding to not holding - between times T_1 and T_2), $declipped(T_1, F, T_2)$ (a fluent F is declipped - from not holding to holding - between times T_1 and T_2), $initially(F)$ (a fluent F holds from the initial time, say time 0), $happens(O, T)$ (an operation O happens at a time T), $initiates(O, T, F)$ (a fluent F starts to hold after an operation O at time T) and $terminates(O, T, F)$ (a fluent F ceases to hold after an operation O at time T). Roughly speaking, in a planning setting the last two predicates represent the cause-effects links between operations and fluents in the modelled world. We will also use a meta-predicate $precondition(O, F)$ (the fluent F is one of the preconditions for the executability of the operation O).

KB_{plan} consists of domain-independent rules and domain-dependent rules. The basic domain-independent rules are

$$
\begin{aligned}
&holds\_at(F, T_2) \leftarrow \\
&\qquad happens(O, T_1) \wedge \\
&\qquad initiates(O, T_1, F) \wedge \\
&\qquad T_1 < T_2 \wedge \\
&\qquad not\, clipped(T_1, F, T_2) \\
&holds\_at(\neg F, T_2) \leftarrow \\
&\qquad happens(O, T_1) \wedge \\
&\qquad terminates(O, T_1, F) \wedge \\
&\qquad T_1 < T_2 \wedge \\
&\qquad not\, declipped(T_1, F, T_2) \\
&holds\_at(F, T) \leftarrow \\
&\qquad initially(F) \wedge \\
&\qquad 0 \leq T \wedge \\
&\qquad not\, clipped(0, F, T) \\
&holds\_at(\neg F, T) \leftarrow \\
&\qquad initially(\neg F) \wedge \\
&\qquad 0 \leq T \wedge \\
&\qquad not\, declipped(0, F, T) \\
&clipped(T_1, F, T_2) \leftarrow \\
&\qquad happens(O,\, T) \wedge \\
&\qquad terminates(O, T, F) \wedge \\
&\qquad T_1 \leq T < T_2 \\
&declipped(T_1, F, T_2) \leftarrow \\
&\qquad happens(O,\, T) \wedge \\
&\qquad initiates(O, T, F) \wedge \\
&\qquad T_1 \leq T < T_2
\end{aligned}
$$

The *domain-dependent rules* define $initiates$, $terminates$, and $initially$, e.g.

$$
\begin{aligned}
&initiates(go(X, L_1, L_2), T, at(X, L_2)) \leftarrow \\
&\qquad holds\_at(mobile(X), T) \\
&initiates(go(X, L_1, L_2), T, free(L_1)) \leftarrow \\
&\qquad holds\_at(mobile(X), T) \wedge \\
&\qquad L_1 \neq L_2 \\
&terminates(go(X, L_1, L_2), T, at(X, L_1) \leftarrow \\
&\qquad holds\_at(mobile(X), T) \wedge \\
&\qquad L_1 \neq L_2 \\
&terminates(go(X, L_1, L_2), T, free(L_2) \leftarrow \\
&\qquad holds\_at(mobile(X), T) \wedge \\
&\qquad L_1 \neq L_2 \\
&initially(at(ag_1, (1, 1)))
\end{aligned}
$$

Namely, the operation go from one location L_1 to some other location L_2 initiates the agent (robot) X being at location L_2 and location L_1 being free and terminates X being at location L_1 and location L_2 being free, provided that X is mobile. Moreover, some agent ag_1 is initially at location $(1,1)$. The conditions for the rules defining $initiates$ and $terminates$ can be seen as preconditions for the effects of the operator go to take place. Preconditions for the executability of operators are specified within KB_{pre}, which contains a set of rules defining the predicate $precondition$, e.g.

$$
\begin{array}{l}
precondition(go(X,L_1,L_2), at(X,L_1)) \\
precondition(go(X,L_1,L_2), free(L_2))
\end{array}
$$

namely the preconditions of the operator $go(X,L_1,L_2)$ are that X is at the initial location L_1 and that location L_2, where X is moving to, is free.

In order to accommodate planning we will assume that the domain-independent part also contains the rules:

$$happens(O,T) \leftarrow assume\_happens(O,T)$$

i.e. actions (by the agent itself) can be made to happen simply by assuming them (by abduction). Note that the abduction of atoms $assume\_happens(a[ag],t)$ by agent ag amounts to planning to execute the corresponding action $a[ag,t]$ by agent ag, to achieve a goal initiated by that action. Similarly, atoms $holds\_at(l,t)$ in the event calculus language correspond to goals $l[t]$ in the state of agents.

The domain-independent part of KB_{plan} also contains the following domain-independent integrity constraints:

$$holds\_at(F,T) \wedge holds\_at(\neg F,T) \Rightarrow false$$

$$
\begin{array}{l}
assume\_happens(O,T) \wedge \\
precondition(O,P) \Rightarrow holds\_at(P,T)
\end{array}
$$

$$
\begin{array}{l}
assume\_happens(O,T) \wedge \\
not\ executed(O,T) \wedge \\
time\_now(T') \Rightarrow T > T'
\end{array}
$$

namely a fluent and its negation cannot hold at the same time, when assuming (planning) that some action will happen, we need to enforce that each of its preconditions hold and that this action will be executable in the future.

To allow agents to draw conclusions from the contents of KB_0, which represent the "narrative" part of the agent's knowledge, the following *bridge rules* are also amongst the domain independent rules of KB_{plan}:

$$
\begin{array}{l}
clipped(T_1,F,T_2) \leftarrow \\
\qquad observed(\neg F[\_],T) \wedge \\
\qquad T_1 \leq T < T_2 \\
declipped(T_1,F,T_2) \leftarrow \\
\qquad observed(F[\_],T) \wedge \\
\qquad T_1 \leq T < T_2 \\
holds\_at(F,T_2) \leftarrow \\
\qquad observed(F[\_],T_1) \wedge \\
\qquad T_1 \leq T_2 \wedge \\
\qquad not\ clipped(T_1,F,T_2) \\
holds\_at(\neg F,T_2) \leftarrow \\
\qquad observed(\neg F[\_],T_1) \wedge \\
\qquad T_1 \leq T_2 \wedge \\
\qquad not\ declipped(T_1,F,T_2) \\
happens(O,T) \leftarrow executed(O,T) \\
happens(O,T) \leftarrow observed(\_,O[T'],T)
\end{array}
$$

Note that we assume that the value of a fluent literal is changed according to observations only from the moment the observations are made, and actions by other agents have effects only from the time observations are made that they have been executed, rather than by the execution time itself. These choices are dictated by the rationale that observations can only be considered and reasoned upon from the moment the planning agent makes them.

Taken KB_{plan} and a goal, the planning capability returns a plan, namely a set of actions and sub-goals, such that by assuming them the goal can be proven to hold in KB_{plan}. For a formal definition, see (Bracciali et al., 2004).

KB_{react} **and Reactivity capability.** The Reactivity capability supports the reasoning capability of reacting to stimuli from the external environment as well as to decisions taken while planning. The capability introduces goals and actions in order to react to some observation recorded in (the KB_0 part of) the given KB or to some goals in $Goals$ and actions in

$Plan$. The reactivity capability allows us to incorporate condition-action rule behaviour, plan repair, policy-based communication. Moreover, as we will see later, this capability allows us to endow agents with norm-conformant behaviour.

The knowledge base KB_{react} represents the knowledge required for reactivity and it is suitably adopted as an extension of the knowledge base KB_{plan}, by adding domain-dependent *reactive rules*, of the form

$$Body \Rightarrow Reaction \wedge TCs$$

where

- $Body$ is a non-empty conjunction of items of the form
  - $observed(l[\_], T)$, where l is a fluent literal,
  - $executed(a[T'], T'')$, where $a[T']$ is a timed action operator,
  - $observed(ag, a[T'], T'')$, where $a[T']$ is a timed action operator and ag is the agent holding KB_{react},
  - $holds\_at(l, T')$, where $l[T']$ is a timed fluent literal,
  - $happens(a, T')$, where $a[T']$ is a timed action operator, and
  - temporal constraints;
- $Reaction$ is either $holds\_at(p, T)$, p being a fluent literal, or $assume\_happens(a, T)$, a being an action, and
- TCs are temporal constraints.

All variables in $Body$ are implicitly universally quantified over the whole implication. All variables in $Reaction \wedge TCs$ not occurring in $Body$ are implicitly existentially quantified on the righthand side of the implication.

Intuitively, a reactive rule $Body \Rightarrow Reaction \wedge TCs$ is to be interpreted as follows: if (some instantiation of) all the observations in $Body$ hold in KB_0 and (some corresponding instantiation of) all the remaining conditions in $Body$ hold, then (the appropriate instantiation of) $Reaction$, with associated (the appropriate instantiation of) the temporal constraints TCs, should be added to $Goals$ or $Plan$ (depending on the nature of $Reaction$).

Taken KB_{react}, the reactive capability returns a set of actions and goals and temporal constraints, such that all the reactive rules in KB_{react} can be proven to hold in KB_{plan} extended with the state resulting from reactivity. The actions will then have to be part of any plan by the agent, and the goals will need to be planned for so that they hold.

3 Extending KGP with Normative Concepts

So far we have summarised the main features of the KGP model for agency. Here we will show how this model can incorporate normative reasoning, simply by adapting the event calculus used originally in the KGP model, for planning and reactivity. The type of scenario that motivates our work is that of an artificial society of agents relying upon an organisation and division of tasks. We will interpret responsibility for tasks of an agent in terms of the roles an agent has been assigned to play in a social environment. Roles are associated with obligations, permission and prohibition, used to define what is expected of the agent when playing a role. Rather than developing or deploying any fully-fledged normative theory, we concentrate on a simple theory (for "responsible" agents) and explore its interaction with the existing operations of KGP agents and the provable conformance of the agents to this theory.

Reactivity is used to make agents aware of their obligations and prohibitions, depending on their observations and the environment in which they are situated. It is also used to force the agents to conform to any relevant normative rules. Planning is used to plan activities that would allow agents to achieve their obligations while avoiding their prohibitions, if possible. As we will see, we define the normative concepts in such a way that agents may have information about other agents' obligations and prohibitions, and thus possibly exploit this information towards the achievement of their objectives.

3.1 Actions and Fluents

In general obligations and prohibitions will be on performing actions or bringing about fluents. Within obligations and prohibitions we will represent actions as terms of the form

$$act(Act, Actor, Parameters).$$

Act names the action, *Actor* is the agent to carry out the act, and *Parameters* is the set of attributes that further specify the action. For instance, the term:

$$act(pay\_fine, ag_1, ''W129FGC'', 60)$$

represents the action of agent ag_1 having to pay a fine of 60 pounds for a car with registration number W129FGC.

For representational convenience, we will often abuse terminology and write $act(Act, Actor)$ if we are only interested in the action and the agent of the action.

We shall represent fluents within obligations and prohibitions as terms of the form:

$$fluent(Fluent, Actor, Parameters).$$

Fluent is the name of the fluent, *Actor* is the agent to bring about *Fluent*, and, as before, *Parameters* represents the rest of the attributes describing the *Fluent*. For instance, the term:

$$fluent(has\_driving\_licence, ag_1,'' ST340578KS'')$$

represents the fluent that agent ag_1 has a driving licence with number ST340578KS.

As before, we will often abuse terminology and write $fluent(Fluent, Actor)$ if we are only interested in the fluent and the agent.

3.2 Roles

In our model the agents' responsability for tasks and assignemnt of obligations and prohibitions is determined by the *roles* the agents play in their social environment. Therefore, we incorporate within the KGP model roles assigned to agents. We shall use the fluent:

$$role(Agent, Role)$$

to state when a specific *Role* is assigned to a specific *Agent*. $role/2$ fluents are initiated and terminated by means of event calculus $initiates/3$ and $terminates/3$ predicates. For instance, the addition of the following domain-specific rule for $initiates/3$:

$$\begin{aligned}&initiates(assign(Y, X, t\_w(W, F, T),\\&\qquad T',\\&\qquad role(X, t\_w(W, F, T)))\end{aligned}$$

allows us to conclude that after the occurrence of an action of the form

$$assign(ag_1, ag_2, t\_w(chelsea, 9, 17))$$

ag_2 plays the role of traffic warden (t_w) in $chelsea$ between times 9 and 17.

Depending on the ontology, we can qualify the assignment of roles, e.g. by means of a rule such as

$$\begin{aligned}&initiates(assign(Y, X, t\_w(W, F, T),\\&\qquad T',\\&\qquad role(X, t\_w(W, F, T))\\&\qquad) \leftarrow\\&authorised(act(assign(Y, X, t\_w(W, F, T), Y), T')\end{aligned}$$

which relies upon a notion of "authorisation". We will not address this notion further in this paper.

Note that roles of agents might change as time progresses and will be initiated or terminated as a result of events happening in the social environment in which an agent is situated.

3.3 Obligations

We represent obligations as atoms of the form:

$$obliged(act(Act, Actor), T, TCs)$$
$$obliged(fluent(Fluent, Actor), T, TCs)$$

We read atoms of the first kind as follows: there is an obligation on the agent *Actor* to bring about the action specified by *Act* (with the appropriate *Parameters*) at a time *T* that satisfies the temporal constraints specified in *TCs*. We read atoms of the second kind as follows: there is an obligation on the agent *Actor* to bring about the fluent *Fluent* to hold at a time *T* that satisfies the temporal constraints specified in *TCs*.

Obligations held by individual agents are generated by means of reactive rules in KB_{react} in the knowledge base of the agents. These reactive rules take the general forms:

$$\begin{aligned}&holds\_at(role(Actor, Role), T_{now}) \wedge\\&Conditions(Actor, T_{now}) \Rightarrow\\&obliged(act(Act, Actor), T_{next}, TCs))\end{aligned}$$

$$\begin{aligned}&holds\_at(role(Actor, Role), T_{now}) \wedge\\&Conditions(Actor, T_{now}) \Rightarrow\\&obliged(fluent(Fluent, Actor), T_{next}, TCs))\end{aligned}$$

The rules above can be read as follows: if an *Actor* is playing a *Role* at time T_{now} and the *Conditions* hold in the knowledge base of the *Actor* at T_{now}, the $Actor$ must fulfill the obligation of making the *Act* happen at T_{next} or bring about the property specified by *Fluent* (with the appropriate *Parameters*) at T_{next}, respectively, with T_{next} satisfying the temporal constraints

TCs. Example instances of the first rule above are given in section 4.

Obligations will be interpreted differently by different agents, depending on their personalities. For example, a “responsible” agent will try to fulfill all its obligations, if possible. We can specify this with reactive rules of the form:

$$\begin{aligned}&holds\_at(responsible(Actor), T) \wedge\\ &obliged(act(Act, Actor), T, TCs) \Rightarrow\\ &happens(Act(Actor), T) \wedge TCs\end{aligned}$$

$$\begin{aligned}&holds\_at(responsible(Actor), T) \wedge\\ &obliged(fluent(Fluent, Actor), T, TCs) \Rightarrow\\ &holds\_at(Fluent(Actor), T) \wedge TCs\end{aligned}$$

referred to as $(R1)$ and $(R2)$, respectively.

3.4 Prohibitions

Prohibitions can be specified in a similar way to obligations. We use atoms of the form

$$\begin{aligned}&prohibited(act(Act, Actor), T, TCs)\\ &prohibited(fluent(Fluent, Actor), T, TCs)\end{aligned}$$

to indicate that $Actor$ is prohibited from performing Act (with the appropriate *Parameters*) at time T within the time constraints TCs, and $Actor$ is prohibited from bringing about $Fluent$ (with the appropriate *Parameters*) to hold at time T within the time constraints TCs, respectively.

Similarly to obligations, prohibitions for individual agents are generated by means of reactive rules in KB_{react} in the knowledge base of the agents. These reactive rules take the general forms:

$$\begin{aligned}&holds\_at(role(Actor, Role), T_{now}) \wedge\\ &Conditions(Actor, T_{now}) \Rightarrow\\ &prohibited(act(Act, Actor), T_{next}, TCs))\end{aligned}$$

$$\begin{aligned}&holds\_at(role(Actor, Role), T_{now}) \wedge\\ &Conditions(Actor, T_{now}) \Rightarrow\\ &prohibited(fluent(Fluent, Actor), T_{next}, TCs))\end{aligned}$$

A concrete example of the first rule is

$$\begin{aligned}&holds\_at(role(X, t\_w(W, F, T), T') \wedge\\ &F \leq T' \leq T \Rightarrow\\ &prohibited(act(leave(W), X), T'', T'' = T')\end{aligned}$$

This rule says that a traffic warden is prohibited to leave his allocated area while he is on duty.
The following rules added to KB_{plan} of the agents then specify the behaviour of responsible agents in the light of prohibitions:

$$\begin{aligned}&holds\_at(responsible(Actor), T) \wedge\\ &prohibited(act(Act, Actor), T, TCs) \wedge\\ &happens(Act(Actor), T) \wedge TCs \Rightarrow false\end{aligned}$$

$$\begin{aligned}&holds\_at(responsible(Actor), T) \wedge\\ &prohibited(fluent(Fluent, Actor), T, TCs) \wedge\\ &holds\_at(Fluent(Actor), T) \wedge TCs \Rightarrow false\end{aligned}$$

These rules indicate that a responsible agent does not perform prohibited acts and does not bring about prohibited fluents to hold.

4 An Example

In this section we exemplify the use of the proposed simple formalisation of normative concepts for “responsible” KGP agents as an executable specification, to which the behaviour of the agents is guaranteed to conform.

Suppose we have two agents, ag_1 and ag_2, with ag_1 a responsible owner of car $W129FGC$ and ag_2 a responsible traffic warden in the $chelsea$ area from time 9 to time 17. They both hold, in addition to rules $(R1)$ and $(R2)$ above, the following reactive rules in their KB_{react}:

$$\begin{aligned}&holds\_at(role(X, t\_w(W, F, T), T') \wedge\\ &observed(parked(C, W), T') \wedge\\ &F < T' < T \Rightarrow\\ &obliged(act(issue\_fine(C), X), T'', T'' = T' + 1))\end{aligned}$$

$$\begin{aligned}&holds\_at(role(X, owner(C)), T) \wedge\\ &observed(issued\_fine(C), T) \Rightarrow\\ &obliged(act(pay\_fine(C), X), T', T' = T + 5))\end{aligned}$$

referred to as $(R3)$ and $(R4)$, respectively.

Assume now that, starting at time 10, agent ag_1 performs three transitions in sequence: Passive Observation Introduction (POI), Reactivity (RE), and

Action Execution (AE). Suppose through the POI, at time 10, agent ag_1 observes that car $W129FGC$ is parked in the $chelsea$ area. POI records this observation in the KB of ag_1 in the form of the atom

$observed(parked(''W129FGC'', chelsea), 10)$

Then RE allows ag_1 to generate for itself the action of issuing a fine to the car at time 11. This is done first by the observation triggering reactive rule $(R3)$ resulting in the derivation of

$obliged(act(issue\_fine(''W129FGC''), ag_1),$
$\quad T,$
$\quad T = 10 + 1)$

which, in turn, triggers reactive rule $(R1)$, resulting in the derivation of

$happens(issue\_fine(ag_1,'' W129FGC''), T) \wedge$
$T = 10 + 1$

The RE transition adds the action to the $Plan$ of agent ag_1. Then the AE transition selects and executes the action by adding a record of its execution to the KB of ag_1.

Now assume that, starting from time 11, ag_2 executes the POI and RE transitions respectively. POI allows ag_2 to observe that it has been issued a fine:

$observed(issued\_fine(''W129FGC''), 11)$

RE then allows ag_2 to generate first an obligation for itself to pay the fine:

$obliged(act(pay\_fine(''W129FGC''), ag_2),$
$\quad T,$
$\quad T = 11 + 5)$

and then an action

$happens(pay\_fine(ag_2, W129FGC), T) \wedge$
$T = 11 + 5$

which will be added to ag_2's $Plan$. If ag_2 then applies the Plan Introduction transition PI next, this may lead to further changes in its $Plan$ and $Goals$. For example in KB_{plan} of ag_2 the preconditions of pay_fine may be specified as follows:

$precondition(pay\_fine(X, Y), have\_money(X))$
$precondition(pay\_fine(X, Y), at(X, police\_station))$

In this case two news subgoals may be added to ag_2's $Goals$, namely to have money and be at the police station at time 16. These would then be further planned for using KB_{plan} and transition PI.

5 Related Work

Computational logic approaches to specify the rules of the interaction amongst social agents is an active subject of research. The existing society model (Alberti et al., 2004a) of PROSOCS, for example, is concerned with the specification of public protocols for communicating agents. This model provides a general language for specifying agent-protocols using the notion of expectations as well as a tool that supports verifiction of compliance for these protocols. Instead, our work presented here seeks to complement what we already have in PROSOCS, so that agents can reason and plan with private obligations, so that we distribute and communicate normative concepts such as roles and obligations to different agents, as it normally happens in a human society. We believe to be able to reuse the mapping between expectations and normative concepts discussed in (Alberti et al., 2004b).

The work described in (Artikis et al., 2002) is also based on computational logic and consists of a theoretical framework for providing executable specifications of particural kinds of multi-agent systems, called open computational societies. Three key components are introduced: a social state, social roles, and social constraints. The specification of these concepts is based on and motivated by the study of legal and social systems, and a representation of deontic concepts in a version of the event calculus is presented.

Our framework differs from that of (Artikis et al., 2002) in that we focus on how to reason about social constraints and roles within a single agent using an abductive interpretation of the event calculus. In other words, we do not take a "bird's eye view" of the interaction between agent but an "agent's view", trying to interpret social constraints from the standpoint of a single agent (based on the specific agent architecture of the KGP model) and further showing how such constraints can be used during agent deliberation.

The IMPACT system (Arisha et al., 1999) incorporates obligations and other deontic notions such as permission and prohibition. In common with us they use abductive logic programs as the representation language. But IMPACT incorporates a more extensive theory of deontic concepts, as well as rules for allowing the utilisation of legacy systems. The KGP model, and its implementation in PROSOCS have different aims compared to IMPACT. We aim at building agents that can plan partially, interleave planning, acting and observations, and adaptability of agents incorporating flexible cycle theories that mimic an agent's personality.

The BOID architecture presented in (Broersen et al., 2001) extends the well known BDI model (Rao and Georgeff, 1995) with obligations, thus giving rise to four main components in representing an agent: beliefs, obligations, intentions and desires. The idea of BOID is to find ways of resolving *internal* conficts within each of the four components and across two or more of the components. In order to do so they define agent types including some well known types in agent theories such as realistic, selfish, social and simple minded agents. The agent types differ in that they give different priorities to the rules for each of the four components. For instance, the simple minded agent gives higher priority to intentions, compared to desires and obligations, whereas a social agent gives higher priority to obligations than desires. The use of priorities is with propositional logic formulae to specify the four components and the agent types.

What we have in common with BOID is that we want to extend our model with the addition of obligations. The existing KGP model already resolves some of the conflicts that they address. For example, if there is a conflict between a belief and a prior intention, which means that an intended action can no longer be executed due to the changes in the environment, the KGP agent will notice this and will give higher priority to the belief than the prior intention, allowing the agent in effect to rectract the intended action and, time permitting, to replan for its goals. The KGP model also includes a notion of priority used in the goal decision capability and the cycle theory that controls the behaviour of the agent.

One of the aims of the KGP model and the extension discussed in this paper, and one of the differences with the work of (Broersen et al., 2001), is to allow the interleaving of concluding obligations, planning to achieve them and other goals and recording and utilising observations. This makes our work closer to the more recent work on BOID, see for instance (Dastani and van der Torre, 2004), by providing an executable specification of the obligations in an abductive logic programming setting. However, our work is not about resolving conflicts between obligations and intentions as that of (Dastani and van der Torre, 2004); we do however plan to study conficts of this kind in future work.

6 Concluding Remarks

We have illustrated how to extend the logical model of agency, known as the KGP model, to support agents with normative concepts. By using obligations and prohibitions as an example, the proposed framework has shown how to specify an agent that can reason using normative concepts and combine them in order to plan and decide which action the agent should perform next.

Unlike approaches that are based on a monolithic tool for checking social interactions, the advantage of this work is that obligations and prohibitions can be interpreted within a social agent and communicated to other agents using the KGP model, thus making the extended model suitable for building multi-agent systems applications whose organisation is based on artificial and open societies of agents. This is an issue for future research.

Future work also involves incorporating permissions, rewards and sanctions in the current model and implementing the extended model in the PROSOCS platform to experiment with concrete applications. Finally, future work involves importing into the KGP model more sophisticated normative theories.

Acknowledgments

The authors would like to thank the anonymous referees for their comments and suggestions on an earlier version of this paper. This work was partially funded by the IST programme of the European Commission, Future and Emerging Technologies under the IST-2001-32530 SOCS project, within the Global Computing proactive initiative.

References

Marco Alberti, Federico Chesani, Marco Gavanelli, Evelina Lamma, Paola Mello, and Paolo Torroni. Compliance verification of agent interaction: a logic-based tool. In Robert Trappl, editor, *Proceedings of the 17th European Meeting on Cybernetics and Systems Research, Vol. II, Symposium "From Agent Theory to Agent Implementation" (AT2AI-4)*, pages 570–575, Vienna, Austria, April 13-16 2004a. Austrian Society for Cybernetic Studies.

Marco Alberti, Evelina Lamma, Marco Gavanelli, Paola Mello, Giovanni Sartor, and Paolo Torroni.

Towards a mapping of deontic logic onto an abductive framework. In *Workshop on Agents and Constraints (Agenti & Vingoli), AI*IA*, 2004b.

K. A. Arisha, F. Ozcan, R. Ross, V. S. Subrahmanian, T. Eiter, and S. Kraus. IMPACT: a Platform for Collaborating Agents. *IEEE Intelligent Systems*, 14(2):64–72, March/April 1999.

A. Artikis and J. Pitt. A formal model of open agent societies. In J. Müller, E. Andre, S. Sen, and C. Frasson, editors, *Proceedings of Conference on Autonomous Agents (AA)*, pages 192–193. ACM Press, 2001.

A. Artikis, J. Pitt, and M. Sergot. Animated specifications of computational societies. In C. Castelfranchi and L. Johnson, editors, *Proceedings of Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1053–1062. ACM Press, 2002.

A. Bracciali, N. Demetriou, U. Endriss, A. Kakas, W. Lu, P. Mancarella, F. Sadri, K. Stathis, G. Terreni, and F. Toni. The KGP Model for Global Computing: Computational Model and Prototype implementation. In *Global Computing*, LNCS. Springer-Verlag, 2004. To appear.

Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert der van Torre. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In Jörg P. Müller, Elisabeth Andre, Sandip Sen, and Claude Frasson, editors, *Proceedings of the Fifth International Conference on Autonomous Agents*, pages 9–16, Montreal, Canada, 2001. ACM Press. URL citeseer.ist.psu.edu/broersen01boid.html.

C. Carabelea, O. Boissier, and C. Castelfranchi. Using social power to enable agents to reason about being part of a group. In *Pre-proceedings of ESAW'04*, Toulouse, October 2004.

Cristiano Castelfranchi, Frank Dignum, Catholijn M. Jonker, and Jan Treur. Deliberative normative agents: Principles and architecture. In *Agent Theories, Architectures, and Languages*, pages 364–378, 1999. URL citeseer.lcs.mit.edu/castelfranchi99deliberative.html.

M. Dastani and L. van der Torre. Programming boid agents: a deliberation language for conflicts between mental attitudes and plans. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS'04)*, 2004.

Frank Dignum. Autonomous agents with norms. *Artificial Intelligence and Law*, 7(1):69–79, 1999. URL citeseer.ist.psu.edu/article/dignum99autonomous.ht

Michael N. Huhns and Munindar P. Singh, editors. *Readings in Agents*. Morgan Kaufmann, San Francisco, CA, USA, 1998.

A.J.I. Jones and M.J. Sergot. On the characterisation of law and computer systems: the normative systems perspective, 1993. URL citeseer.ist.psu.edu/jones93characterisation.html. Deontic Logic in Computer Science: Normative System Specification. John Wiley and Sons, Chicester (1993) 275–307.

A.J.I. Jones and M.J. Sergot. A formal characterisation of institutionalised power. *Journal of the IGPL*, 4(3):429–445, June 1996.

A. Kakas, P. Mancarella, F. Sadri, K. Stathis, and F. Toni. Declarative agent control. In Leite J. and Torroni P., editors, *Proceedings CLIMA'04, 5th International Workshop on Computational Logic in Multi-Agent Systems*, Lisbon, Portugal, Sep. 2004a.

A. Kakas, P. Mancarella, F. Sadri, K. Stathis, and F. Toni. The KGP model of agency. In *European Conference on Artificial Intelligence (ECAI04)*, pages 33–39, 2004b.

R. A. Kowalski and M. Sergot. A logic-based calculus of events. *New Generation Computing*, 4(1): 67–95, 1986.

Julian A. Padget, editor. *Collaboration between human and Artificial Societies: Coordination and Agent-based Distributed Computing*. Springer, LNAI 1624, 2001.

A. S. Rao and M. P. Georgeff. BDI-agents: from theory to practice. In *Proceedings of the First Intl. Conference on Multiagent Systems*, San Francisco, 1995. URL citeseer.ist.psu.edu/rao95bdi.html.

K. Stathis, A. Kakas, W. Lu, N. Demetriou, U. Endriss, and A. Bracciali. PROSOCS: a platform for programming software agents in computational logic. In J. Müller and P. Petta, editors, *Proceedings of the Fourth International Symposium "From Agent Theory to Agent Implementation"*, Vienna, Austria, April 13-16 2004.

F. Toni and K. Stathis. Access-as-you-need: a computational logic framework for flexible resource access in artificial societies. In *Proceedings of the Third International Workshop on Engineering Societies in the Agents World (ESAW'02)*, Lecture Notes in Artificial Intelligence. Springer-Verlag, 2002.

On the Potential of Norm-Governed Behavior in Different Categories of Artificial Societies

Paul Davidsson
Department of Systems and Software Engineering
Blekinge Institute of Technology, 372 25 Ronneby, Sweden
paul.davidsson@bth.se

Stefan Johansson
Department of Systems and Software Engineering
Blekinge Institute of Technology, 372 25 Ronneby, Sweden
stefan.johansson@bth.se

Abstract

Based on a classification of artificial societies and the identification of four different types of stakeholders in such societies, we investigate the potential of norm-governed behavior in different types of artificial societies. The basis of the analysis are the preferences of the stakeholders and how they influence the state of the society. A general conclusion drawn is that the more open a society is the more it has to rely on agent owners and designers to achieve norm-governed behavior, whereas in more closed societies the environment designers and owners may control the degree of norm -governed behavior.

1 Introduction

A collection of software entities interacting with each other for some purpose, possibly in accordance with common norms and rules, may be regarded as an *artificial society*. This use of the term "society" is analogous to human and ecological societies. The role of a society is to allow the members of the society to coexist in a shared environment and pursue their respective goals in the presence of others. As a software entity typically acts on the behalf of a person or an institution, i.e., its owner, we will here refer to these entities as *agents*. This use of the term is somewhat more general than is common. However, since the principles we will discuss are general, covering all different kinds of (semi-)autonomous software processes, there is no reason for limiting the discussion to "proper" software agents.

Based on a classification of artificial societies and the identification of four different types of *stakeholders* in such societies, we will investigate the potential of norm-governed behavior in different types of artificial societies. The basis of the analysis are the *preferences* of the stakeholders and how they influence the state of the society.

In the next section, we present the view of artificial society used in this work and the type of stakeholders involved. This is followed by the categorization of agent societies and a discussion of the role of preferences. We then analyze the potential for norm-guided behavior in the different categories based on preferences of the stakeholders and draw a number of conclusions.

2 Agent Societies and Stakeholders

There are a number of other notions used to refer to organizational structures of software agents, e.g., *groups*, *teams*, *coalitions*, and *institutions* (cf. for instance (Carley and Gasser, 1999), (Ferber and Gutknecht, 1998), and (Singh et al., 1999)). Since a society may contain any number of institutions, coalitions, teams, groups, and individual agents, the concept of society belongs to a higher organizational level than these structures. Also, whereas a society is neutral with respect to co-operation and competition, coalitions are formed with the explicit intention of co-operation. Similarly, a team is a group in which the agents have a common goal. The difference between a group or a team and an institution is that an institution has a legal standing distinct from that of individual agents.

Artikis and Pitt (2001) provide a formal characterization of an agent society that includes the following entities:

- a set of agents,
- a set of constraints on the society,
- a communication language,
- a set of roles that the agents can play,

- a set of states of affairs that hold at each time at the society, and
- a set of owners (of the agents).

They describe the set of constraints as "constraints on the agent communication, on the agent behavior that results from the social roles they occupy, and on the agent behavior in general." Another way describing the set of constraints is that they constitute the norms and rules that the agents in the society are supposed to abide. When appropriate, we will refer to the above list of entities when discussing different types of societies.

A limitation of this characterization is that only consider one type of stakeholder of the society, namely the agent owners. Following Johansson (2002), we will here regard three other entities, namely the owner of the society, or rather the *environment* (e.g., computational platform), in which the agents act, and the *designers* of the agents as well as of the environment. By environment *owner* we mean, the person or organization that have the power to decide which agents may enter the society, which roles they are allowed to occupy, what communication language should be used, and the set of constraints on the society. (An alternative characterization of a society would be to include also the computational platform.) More details concerning the stakeholders will be presented in a later chapter.

In the next chapter we will present a categorization of artificial societies based on their degree of openness and discuss their strengths and weaknesses in terms of flexibility, stability and trustfulness.

3 Categories of Agent Societies

In previous work (Davidsson, 2001), four basic types of artificial societies were identified. These are briefly described below, starting with the most straightforward types.

3.1 Open Societies

In principle, it is possible for anyone to contribute one or more agents to an open society without restrictions. An agent joins the society simply by starting to interact with some of the agents of the society.

An society supports openness and flexibility very well, but it is very difficult to make such a society stable and trustful. For instance, it is not possible to control the set of constraints or monitor whether the agents abide these. In fact, it is not possible to determine the set of agents in any effective way. Within an open society, the only structure is typically just a generally accepted low-level communication language and a limited set of generic roles.

The most obvious example of an open artificial society is the World Wide Web (WWW), where the set of members of the society consists of the set of WWW-browser processes together with the set of WWW-server processes that are connected to the Internet. HTTP (Hypertext Transfer Protocol) is the low-level communication language. The number of roles is limited to generic clients, i.e., the browsers, and servers. Finally, the set of owners is either the owners of the machines on which the browser and server processes are run, or the persons/institutions that started the browser and server processes. The openness of the society is obvious in this case; anyone with an Internet connection is allowed to start a browser process or a server process and join the artificial society defined by the WWW without any restrictions. Although there is some kind of environment owner involved, i.e., World Wide Web Consortium (W3C), we will not regard it as such since it cannot impose society-level constraints, roles, and communication language.

3.2 Closed Societies

Closed agent societies are typically those where a Multi-Agent System (MAS) approach is adopted by a team of software developers to implement a complex software system. The MAS is designed to solve a set of problems, specified by the society owner. The solving of these problems is then distributed between the agents of the MAS. It is not possible for an "external agent" to join the society. Zambonelli et al. (2001) refer to this type of systems as "distributed problem solving systems" and describe them as "systems in which the component agents are explicitly designed to co-operatively achieve a given goal." They argue that the set of agents is known a priori, and all agents are supposed to be benevolent to each other and, therefore, they can trust one another during interactions. In open systems, on the other hand, agents are "not necessarily co-designed to share a common goal" and cannot be assumed to be benevolent.

The concept of closed agent society corresponds to the large majority of existing MAS. An advantage of closed societies is that it is possible to precisely engineer the society, e.g., specify exactly which agents interact, and why etc. Consequently, closed societies provide strong support for stability and trustfulness. However, they are able to provide very little openness and flexibility.

3.3 Semi-Open Societies

In semi-open artificial societies, there is an institution that functions as a gate-keeper. For instance, agents wanting to join the society may contact the institution to whom it promises to follow the set of constraints of the society. The institution then makes an assessment whether the agent is trustworthy and eligible and decides whether to let it join the society or not. It is, of course, possible to differentiate between classes of trustworthiness so that agents that are considered more trustworthy are given access to more services etc than agents considered less trustworthy.

In fact, there are already a number of distributed information systems that resemble semi-open societies. For example, consider peer-to-peer systems (Oram, 2001), such as the Internet-based Kazaa service (cf. www.kazaa.com) which let users share their achieves of media files. Each user must use a particular type software (agent), that is either provided by Kazaa or by others given that it follows the same protocols as the original Kazaa agents. In order to get access to other users' files, the agent needs to connect to a central server, which then may let the process join the society. If the process succeeds to join the society, it will be able to interact with other users’ agents, downloading and uploading media files. Thus, anyone may potentially contribute an agent (or more) to the society, but before it joins the society it is registered at the central server.

Semi-open societies only slightly limits the openness compared to completely open societies, but have a much larger potential for providing stability and trustfulness. For instance, it is possible to monitor which agents are currently in the society. This also makes the boundary of the society explicit.

3.4 Semi-Closed Societies

In what we will refer to as semi-closed societies, external agents are not allowed to enter. However, they have the possibility to initiate a new agent in the society which will act on the behalf of the external agent. This is done by contacting some kind of institution representing the society and ask for the creation of an agent (of a predefined type). The institution then creates an agent of this type, which enters the society with the goal of achieving the goals defined by the external agent. See Fig. 1

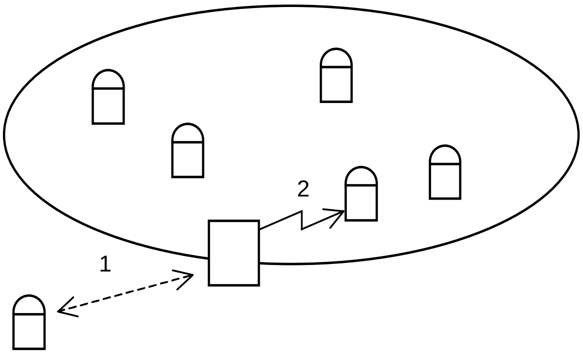


Fig 1: An agent initiating a representative in a semi-closed society

All agents are run on the same (set of) servers. Typically, the agents are implemented and the servers are managed by a third party, i.e., the owner of the environment. As the possible behaviors of all agents are known, it is easier to control the activities in the society.

The following example of a semi-closed society based on the activity of searching for and booking “last-minute” holiday travel tickets. The price of this type of tickets may change on a daily basis and are determined by current supply and demand. Today, customers find and buy tickets manually by browsing newspaper ads and WWW-pages, as well as phoning and visiting local travel agents. Many people regard this as a time-consuming and boring activity and would probably be happy to let software agents do the job. The society is based on an existing prototype implementation further described in Davidsson (2001). Customers specify their preferences (departure date, destination, max. price etc.) through either a WAP or WWW interface. An agent is then initiated on a service portal (at a remote computer) with the goal of finding a ticket satisfying the customer preferences. It continually searches a number of databases until it reaches its goal (or is terminated by the customer). When a ticket is found, the agent either books it directly, or sends an SMS to the customer asking for confirmation. To book the ticket, the customer agent contacts an agent representing the travel company (the info in the database contains the address to the travel agent). When the customer agent receives a confirmation, it immediately sends an SMS message to the customer about this and then terminates.

Semi-closed societies provide almost the same degree of openness as semi-open societies but are less flexible. On the other hand, they have a larger potential for implementing stability and trustfulness. An interesting property of the semi-closed societies is that they seem to indicate the limit of how open a society could be where the owner of the society may still control the overall architecture of the society. To have control over the architecture is a prerequisite for applying many of the ideas on how to achieve multi-agent coordination (Lesser, 1998). Moreover, this type of society poses interesting questions regarding ownership: Who is actually responsible for the actions of the agents?

3.5 Summary of Society Categories

We have described four different categories of artificial societies that balance the trade-off between openness, flexibility, stability, and trustfulness differently. Open societies support openness and flexibility but not stability and trustfulness and the opposite is true for closed societies. Two other, intermediate categories, namely semi-open and semi-closed societies, tries to achieve the best from both worlds. Whereas semi-open societies are more flexible than semi-closed societies, they have lower potential to achieve stability and trustfulness. A summary comparison between the different types of societies (ordered in degree of openness) is provided in Table 1.

Table 1. A comparison between the different types of artificial societies.

	Closed	*Semi-closed*	*Semi-open*	*Open*
Agents	Known at design time	Known at run time	Known at run time	Cannot be known
Const-raints	Fixed	Fixed	Fixed	Not fixed
Commun. language	Fixed	Fixed	Fixed	Not fixed
Roles	Fixed	Fixed	Usually fixed	Not fixed
State can be monitored	Yes	Yes	Partially	No
Agent owners	Fixed at design time	Can be known	Can be known	Cannot be known
Environment owner	Yes	Yes	Yes	No

The balancing of trade-off provided by the semi-open and semi-closed societies is necessary for implementing systems that can be characterized as information ecosystems. In this type of systems there is a strong need for mechanisms for “enforcing” normative behavior between agents in order to provide trustful systems to end-users. In completely open societies such mechanisms probably need to be very complex (if they exist at all), which means that the potential for achieving trustful systems is very low. In completely closed systems, on the other hand, the potential for achieving trustfulness is great, but the price you have to pay, by making it impossible for new agents/owners to enter the society, is too big in applications where openness is desired.

4 Preferences

The behaviors of agents at the micro-level as well as the behavior of the society of agents at the macro-level are to a large extent determined by the *preferences* (Johansson and Kummeneje, 2001) of the stakeholders involved. By preference we here mean an ordering of possible states in terms of their goodness from the view of a stakeholder.

Johansson (2002) has identified four types of stakeholders:

- The agent owner (AO) has the power to launch the agent, provide it with preferences, as well as make run-time decisions regarding updating of preferences and when the agent should be terminated.
- The agent designer (AD) has designed (and possibly implemented) the action selection and execution mechanisms of the agent.
- The environment owner (EO) has the power to launch the environment, provide it with preferences (which are reflected in the dynamics of the rules and the conditions under which agents are able to act in the environment), as well as make run-time decisions regarding updating of preferences and when the environment should be terminated.
- The environment designer (ED) has designed and possibly implemented the rules and conditions under which agents act in the environment.

Let us return to the example of file sharing to illustrate these stakeholders. In the case of Kazaa, Sharman Networks is the EO, i.e. they own the environment originally designed by Zennström and Friis in their company Consumer Empowerment (ED). There are several ADs that design and implement agents. Sherman Networks is one of them, but there are several others, e.g. Kazaa Lite K++ provided by other parties.[1] The users of Kazaa are the AOs, expressing their preferences to their agent about which files to download, which files to share with others, etc.

The preferences of the different stakeholders have different characteristics. For instance, the preferences of the owners are *dynamic* in the sense that they can be changed at run-time, while the preferences of the designers are *static* (non-changeable) and must be set before the agent or environment is launched. Both of these preference types may have an impact on all actions. Even though the designer preferences are static, they play a crucial role deciding which actions the owner can influence via its preferences. Only the preferences that

1. the owner is able to express, and
2. the designer's interpreter is able to interpret

can be adopted by the agent or environment. It is worth noting that it is only the preferences of the AD (and to some extent the ED) that directly influence the behavior of the society. The preferences of the other stakeholders only indirectly the behavior in the sense that they must be mediated by interpreters implemented by the ADs and ED.

5 Norms, Preferences and Artificial Societies

We will now take a closer look at the implications on norm-guided behavior for the different types of artificial societies of the stakeholders and their preferences. According to Artikis and Pitt (2001), norms are part of the “set of constraints on the society”. In this sense, norms can be seen primarily as a tool to meet the preferences of the ED and EO,

[1] Since Sharman Networks business idea is to sell advertisements to the users of the file sharing network by bundling the agent with *malware* (Skoudis and Zeltser, 2003) such as adware, they are trying to prevent the use of agents free from such malware, e.g., by shutting down sites that distribute them referring to the copyright violation. However, these efforts have so far not been very successful.

at least in societies that are not more than semi-open.

5.1 The Preferences of the Society Types

In closed societies, the ED and the ADs are often the same (group of) person(s). It is typically some sort of distributed problem solving system designed to be efficient, rather than being flexible with respect to external changes. Although the society as a whole may be used by others for solving problems at the system level, e.g. to play chess (Drogoul, 1995), the AOs are not able to change the way the pieces move and the EO is not allowed to introduce new agents into the game (unless stated so in the design, e.g. when a peasant reach the opposite side).

In semi-closed societies, the AOs express a number of preferences to the EO, which in turn, create and launch an agent into the society. The interpretations of these preferences are totally in the hands of the AD/ED, but if misused, the AO is not likely to return. In addition, the ED and AD are typically the same, which means that all static preferences can be controlled by the same (group of) person(s) or organization.

In semi-closed as well as semi-open societies, there will have to be some (predicted future) surplus value getting back to the AO from the society, or the AO will not participate and reveal any preferences. One good example of this is the malware that the AD includes in the file sharing agents, to which the AO can express its preferences concerning what files to download. The EO may be able to keep track of what files are downloaded as well as other facts about the AO which then may be distributed to third parties through the malware. If these leakages of information become integrity and security threats, the AO may remove its agent from the society.

In semi-open societies, the AO have better opportunities to keep its preferences within the agent, thus only having to reveal them for the internal interpretation of the agent. The ED may, by providing means of jurisdiction, help the EO to remove those agents that do not meet its preferences.

In open societies it is often very difficult to enforce stakeholders' preferences. The ED can create protocols, platforms etc. that the agents may use, but the agents may also choose not to use them, thus being out of control of the EO who may have a hard time enforce punishments, such as throwing out a misbehaving agent. Instead, the agents within the society may cooperate in order to build trustful coalitions (if this is a preference of the ADs and AOs). Previous work by Johansson (2002) indicates that rational agents are unable to build coalitions unless they are designed to do so. The emergence of coalitions is simply not compatible with the possibility of exploiting the coalition; instead the equilibrium stays in a non-cooperative state. Thus, there have to be explicit AD (and AO) preferences for being cooperative in order to get rational agents to build coalitions.

5.2 The Norms of the Society Types

Norms become increasingly important as we move along the scale from closed to open systems. In semi-open systems, there is support for explicit norm handling implemented by the ED and instantiated by the EO. In addition, the agent themselves may create and join coalitions, and by doing so, increase the number of norms to follow.

Open societies, probably need the norms the most. Paradoxically, in such societies norms often are hard to express in terms of preferences of the stakeholders. We have to rely on the emergence of norms systems, and as mentioned above in the discussion of coalitions, this put a lot of burden on the ADs and AOs. Also, the lack of a given structure of the society leads to problems in creating jurisdictional instruments, unless some sort of agreement is set on the designers level. But as these punishments are realized in the environment, we move towards dealing with a semi-open society rather than an open one.

5.3 The Norms of the Stakeholders

There is a clear distinction between the norms that are designed by the ED (and possibly instantiated by the EO) and the norms that are designed by the AD (and possibly instantiated by the AO). In the former case, the norms are there to improve the value of the society as a whole. These environment norms may be both constitutional, e.g. as in the case of what agents that are allowed to enter a semi-open society, or regulative, as in punishments for misbehaving agents (Boella and Torre, 2004). In the latter case, the norms initiated by the agents in the system, are there in order to improve the situation for the individual agent, e.g., by creating a coalition with other agents. These norms may also be of both kinds. Constitutional norms are necessary in open systems in order to make agent communication efficient, thus these norms (e.g., using a certain ACL) are set by the AD. Regulative norms are present in form of the interpretation (implemented by the AD) of the preferences of the AO.

We may also discuss norms from the designer/owner perspective. Designers of agents and environments will also have to design constitutional norms patterns for their respective implementations. That is, if an agent is going to be able to discuss, create or break regulative norms, they have to be conceptualized by the design (and the same holds for the environment). The contents of these norms can however be filled by their owners (as preferences) or possibly be created by the agents themselves in runtime, given that they are sufficiently autonomous, i.e. norm autonomous (Verhagen, 2000).

6 Future work

Our plan is to develop guidelines for agent and environment designers based on the analysis above. Below we outline such guidelines and discuss some limitations.

The task of the ED is to implement the conditions under which the agents of the society will be run. One way of handling this is to set the rewards and punishments for certain behaviors in a way that will lead the expected behavior to an acceptable behavioral equilibrium.[2] We therefore suggest the following schematic guidelines for environment design (and maintenance):

1. Set up the conditions under which the agents are allowed to act in the society (i.e. formulate the constitutional norms).
2. Assign preferences to each (class of) possible state(s), describing the estimated value of these state(s) (from the perspective of the ED-EO).
3. Calculate the assignment of punishments and rewards of behaviors that, when implemented in the environment as regulative norms, will bring the society to an equilibrium in the preferred states.

We may expect a further development of the skills and abilities of the agents as the field of agent engineering matures. This means that they will be able to (if possible) exploit the weaknesses of the environments that they act in, as well as that of other agents, i.e. an *arms race* (Carlsson, 2001). Today these weaknesses are exploited manually through the expression of explicit owner preferences, but as the level of abstraction increases, we may expect this to be automated in a way.

A suggested set of guidelines for ADs are therefore to design and implement:

1. Abilities to find out what rules and conditions that are applied in the environment (e.g. by asking look-up services, etc).
2. Abilities to optimize the behavior with respect to:
    a. the actions possible to perform in the given environment, i.e. the constitutional norms
    b. the expected rewards and punishments of different behaviors in the society, (i.e. the regulative norms of the environment and the norms of agent coalitions), and
    c. the preferences of the AO.
3. An interface to the AO in which it can express its preferences.

6.1 Type specific limitations

The method above is not general in the sense that it is applicable in all types of agent societies. In the case of open societies, the ED may design certain constitutional norms, such as a basic communication protocol, or a common platform, but when it comes to regulative norms, it gets more troublesome, since the EO lacks the jurisdiction to take care of misbehaving agents. This task is instead adopted by the agent owners and may result in e.g. a closed connection in the case of the WWW. The agents are also (in the best of worlds) able to discuss, create and maintain norms within coalitions in the open society.

In semi-open societies, the gatekeeper will have the ability to throw out agents that do not live up to the norms of the environment, thus constitutional as well as regulative norms are easier to maintain than in an open society. Still, the agents are designed and spawned outside the environment, making them capable of creating coalitions without the intervention of the environment.

In semi-closed and closed societies, however, this is not the case. The agents are created within the system and the AO has a very limited possibility to let the agents create coalitions unless that opportunity is stated in the AO preferences given when the agent is to enter the system. These type of systems have a good support for constitutional norms set by the EO.

7 Conclusions

We identify two types of norms; the static constitutional norms set by the designers of the environment and the agents, and the (possibly dynamic) regulative norms set by the designers and the owners jointly where the designers may implement the norm, or implement a template that is instantiated at run time by the owners when they express their preferences to the system.

Different types of agent societies are able to control the state of the society to different extent. In open societies the agent preferences (both designer and owner) will completely decide what will happen in the society. However, the more closed the society is, the larger is the potential for using environment preferences to influence the state of the society. Thus, if norms are viewed as environment preferences, different types of agent societies support norm-guided behavior to different extent.

References

Artikis, A. and J. Pitt, 2001. A Formal Model of Open Agent Societies, *Proceedings of the Fifth*

[2] By the term behavioral equilibrium, we mean that there is a balance in the system of punishments and rewards in the sense that the choice of behaviors (at some level of granularity) is stable.

International Conference on Autonomous Agents.

Boella, G. and L. v d Torre, 2004. Regulative and Constitutive Norms in Normative Multiagent Systems. *Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning*, AAAI Press, p.255-266.

Carley, K.M. and L. Gasser, 1999. Computational Organization Theory. In G. Weiss (editor), *Multiagent Systems*, MIT Press.

Carlsson, B., 2001. Conflicts in Information Ecosystems, - Modeling selfish agents and antagonistic groups, Ph.D. diss., Department of Software Engineering and Computer Science, Blekinge Institute of Technology, Sweden.

Davidsson, P., 2001. Categories of Artificial Societies, In *Engineering Societies in the Agents World II*, Springer Verlag LNCS series, Vol. 2203.

Drogoul, A., 1995. When Ants Play Chess (Or Can Strategies Emerge from Tactical Behaviors), *Proceedings of MAMAAW 93*, Springer Verlag LNAI series, Vol. 957.

Ferber, J. and O. Gutknecht, 1998. A Meta-model for the Analysis and Design of Organizations in Multi-Agent Systems. *Proceedings of the Third International Conference on Multi-Agent Systems*, IEEE Computer Society.

Johansson, S.J., 2002. On Coordination in Multi-Agent Systems, Ph.D. diss., Department of Software Engineering and Computer Science, Blekinge Institute of Technology, Sweden.

Johansson, S.J. and J. Kummeneje, 2001. A Preference-Driven Approach to Agent Systems. *Proceedings of the Second International Conference on Intelligent Agent Technologies*.

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*, Vol. 1, pp. 89-111, Kluwer.

Oram, A. (editor), 2001. *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*, O'Reilly.

Singh, M.P., A.S. Rao, and M.P. Georgeff, 1999. Formal Models in DAI, In G. Weiss (editor), *Multiagent Systems*, MIT Press.

Skoudis, E., and L. Zeltser, 2003. *Malware – Fighting Malicious Code*, Prentice Hall, first edition.

Verhagen, H., 2000. Norm Autonomous Agents, Ph.D. diss., Department of Computer and System Sciences, Royal Institute of Technology, Sweden.

Zambonelli, F., N.R. Jennings, and M. Wooldridge, 2001. Organizational Abstractions for the Analysis and Design of Multi-Agent Systems. In *Agent-Oriented Software Engineering*. Springer Verlag LNCS series, Vol. 1957.

Ontological Aspects of the Implementation of Norms in Agent-Based Electronic Institutions

Davide Grossi* davide@cs.uu.nl
Huib Aldewereld* huib@cs.uu.nl
Javier Vázquez-Salceda* javier@cs.uu.nl
Frank Dignum* dignum@cs.uu.nl

*Institute of Information and Computing Sciences
Utrecht University
PO Box 80.089 3508TB Utrecht
The Netherlands

Abstract

In order to regulate different circumstances over an extensive period of time, norms in institutions are stated in a vague and often ambiguous manner, thereby abstracting from concrete aspects, which are relevant for the operationalisation of institutions. If agent-based electronic institutions, which adhere to a set of abstract requirements, are to be built, how can those requirements be translated into more concrete constraints, the impact of which can be described directly in the institution? We address this issue considering institutions as normative systems based on articulate ontologies of the agent domain they regulate. Ontologies, we hold, are used by institutions to relate the abstract concepts in which their norms are formulated, to their concrete application domain. In this view, different institutions can implement the same set of norms in different ways as far as they presuppose divergent ontologies of the concepts in which that set of norms is formulated. In this paper we analyse this phenomenon introducing a notion of contextual ontology. We will focus on the formal machinery necessary to characterise it as well.

1 Introduction

Electronic institutions (eInstitutions) are agent environments that can regulate and direct the interactions between agents, creating a safe and stable environment for agents to act. This is accomplished by incorporating a number of norms in the institution which indicate the type of behaviour to which each agent should adhere within that institution. Similar to their human counterparts (legal systems are the eminent example), norms in eInstitutions should be stated in such a form that allows them to regulate a wide range of situations over time without need for modification. To guarantee this stability, the formulation of norms needs to abstract from a variety of concrete aspects, which are instead relevant for the actual implementation of an eInstitution, see for instance Dignum (2002) and Grossi and Dignum (2004); this means that norms are expressed in terms of concepts that are, on purpose, kept vague and ambiguous, cf. Hart (1961). On the other hand, whether a concrete situation actually falls under the scope of application of a norm is a question that, from the point of view of an effective operationalisation of the institution, should be answered in a clear and definite way.

The problem is that concrete situations are generally described in terms of ontologies which differ from the abstract ontology in which, instead, norms are specified. This means that, to actually give a concrete operational meaning to the norms, i.e., to implement them, a connection should be made which can integrate the two ontological levels as sketched in Dignum (2002). We need to determine what the concepts in the situation mean and somehow check them against the terms used in the norms. In other words, we have to see whether the concepts used to specify the situation are classified by (or counts as) the concepts used in the norm formulations; we have to formulate them in an ontology which makes the relation between the concrete and the abstract specifications explicit.

In previous work we have focused on declarative aspects of norms, see Dignum et al. (2002) and Dignum et al. (Oct. 2002), formally defining norms by means of some variations of deontic logic that include conditional and temporal aspects, in Broersen et al. (2004) and Dignum et al. (2004). We have also explored some of the operational aspects of norms, by focusing on how norms should be operationally implemented in MAS from an institutional perspective in Vázquez-Salceda et al. (2004). In this paper we extend this line of research, taking into account the ontological aspects of norm implementation.

This work is organised as follows. In the next section we will elaborate on how ontologies are used in institutions to determine the meaning of the concepts used in the norms under different contexts. Then, in section 3, we will present a formal framework in which it is possible to represent and reason about divergent ontologies (we will also call them contextual terminologies) based on Grossi et al. (2004a,b). Using this framework we will formalise an example in section 4. After this, we will discuss the implementational aspects of our framework in section 5 and we end the paper with some discussion, conclusions and future work.

Throughout the paper, we will use the regulations on personal data protection in several scenarios: the European Union, the Dutch Police, European Hospitals and the Spanish National Transplants Organisation (an organisation for the allocation of human organs and tissues for transplantation purposes).

2 Institutions, Ontologies and Contexts

In order to properly implement norms in eInstitutions, we should first analyse how norms are handled in human institutions. It is our thesis that institutions provide structured interpretations of the concepts in which norms are stated. In a nutshell, institutions do not only consist of norms, but are also based on ontologies of the to-be-regulated domain. For instance, whether something within a given institution counts as *personal data* and should be treated as such depends on how that institution interprets the term `personal_data`. What counts as personal data in a hospital, might not count as personal data in a police register and vice versa. Nevertheless, in both hospitals and police registers, if some piece of information is personal data, it should then be treated in accordance to the regional, national and/or international privacy policies. That is to say, hospitals and police registers, although providing potentially inconsistent understanding of what personal data is, do share the normative consequences (rights, duties, prohibitions, etc.) attached to the classification of information as personal data.

This perspective on institutions, which emphasises the semantic dependence of norm implementation, goes hand in hand with widely acknowledged positions on the normative nature of social reality. Institutions can be indeed seen as normative systems of high complexity, which consist of regulative as well as non-regulative components (see Alchourrón and Bulygin (1986), Jones and Sergot (1993), Jones and Sergot (1992), Searle (1995) and Boella and Van der Torre (2004)), that is to say, which do not only regulate existing forms of behaviour, but they actually specify and create -via classification- new forms of behaviour. In legal theory, the non-regulative component of the issuing of norms has been labelled in ways that emphasise a classificatory, as opposed to a normative/regulative, character: *determinative rules* (Von Wright (1963)), *conceptual rules* (Bulygin (1992)), *qualification norms* (Peczenik (1989)), *definitional norms* (Jones and Sergot (1992)). This characteristic of the non-regulative, or classificatory, components of normative systems is intermingled with a second feature, namely the *constitutive*, *conventional* character of these components that have therefore been called also *constitutive rules* or *constitutive norms*, cf. Ross (1968) and Searle (1995). In this view, statements to the effect that racial data count as personal data establish that being racial data constitutes, in the sense of being a sufficient condition, for being personal data. However, this "constitution" is not absolute. It being conventional, it only holds within the specific institution in which that relation of constitution is effective, it is *contextual*. This feature has been particularly emphasised in Searle (1995), where constitutive rules are viewed as representable via the following type of statements: "X counts as Y *in context* C".

2.1 Context

Human institutions hardly operate in isolation and therefore frequent references are made to other regulations and institutions. Institutions and their environment are interdependendent, and each influences the other. In human societies the context of an institution includes regulations that are applied to the institution's internal and/or external behaviour. Therefore, when building eInstitutions, special attention should be given to the environment where the eInstitution will operate, cf. Vázquez-Salceda (2004), as the environment may affect its specification (especially in the normative aspects of the eInstitution) and design;

the regulations that apply to the environment should be considered and included by the designer inside the designing process of the eInstitution.

In agent-based eInstitutions, the agents should be provided with a model of the norms that may apply inside the institution and an ontology giving an interpretation of the terms used. From the point of view of a single eInstitution, a single norm model and ontology are enough in order to define the boundaries between acceptable and unacceptable behaviour. But problems may arise when agents have to operate in more than one eInstitution, each one having its own norms and norm interpretation, or when two eInstitutions have to inter-operate. The source of these problems is that, in most real domains, norms are not universally valid but bounded to a given *context*. This is the case of norms, for instance, in Health Care, as they are bounded to transnational, national and regional regulations, each of them defining a different normative context.

In those scenarios where more than one normative context should be modelled trying to force a single vocabulary, theory and representation to model and reason about any situation on any context is not a good option. The alternative, first proposed by McCarthy in McCarthy (1986, 1987), is to include *contexts* as formal objects in the model. Therefore, most theoretical approaches have moved towards having an explicit representation of context. One of the most used approaches is the *box metaphor*, that is, considering context as a box (from Giunchiglia and Bouquet (1997)):

> *[...] Each box has its own laws and draws a sort of boundary between what is in and what is out.*

With this idea, in Vázquez-Salceda (2004), context in eInstitutions is defined formally as a subset of possible worlds where there is a shared vocabulary and a normative framework to be followed by a certain group of agents. In this view, an eInstitution is a context defining a) its vocabulary (by means of an ontology) and b) the norms that apply in that context. In parallel, the environments where the eInstitution operates are also (super)contexts, being possibly nested (e.g. to model the nesting in regional/national/transnational environments).

2.2 Contextual Ontologies

Each normative context should therefore define a vocabulary to be shared by agents in that context. It means that each context is associated with a domain ontology that defines the meaning of the terms that are present in the norms, the actions the agent may perform and the terms in the communication with others. However, standard ontologies are not enough. As we have mentioned, contexts may be nested. Each context (defining their norms and an ontology) may contain other (sub)contexts inside (extending and/or modifying the norms and the ontology) or belong to one or several (super)contexts. Some kind of connection should be made between ontologies of interrelated contexts. This problem usually appears in multiagent systems that should operate in a transnational, multi-lingual environment such as Europe. To illustrate this problem, let us return to the regulations on personal data protection. In European Union regulations[1] *personal data* are defined as *"[...] those [data] which allow the identification of a person, and which reveal racial or ethnic origins, political opinions, religious or philosophic beliefs, trade union's affiliation, as well as data related to health or sexuality"*. This abstract definition of the term *personal data* has been introduced, in more or less extent, in the regulations of the EU member states. EU regulations on personal data protection apply to every data archives structured in a way which allows the easy extraction of personal information, including electronic archives on any computer-readable storage device and format. One important aspect is the rights that EU citizens have over their personal data:

- *individual's consent*: as a general rule, personal data collection and processing requires the approval of the affected person.

- *rights over the collected information*: each person has the right to access, amend, cancel or be opposed to the collection of her personal data,

- *data maintenance*: Personal data will only be kept during the period needed to achieve the aims they were collected for, or the authorised extensions of those aims. If it is desirable to maintain this information long after this period (for historical, statistical or scientific purposes), it must be done in a way that avoids personal identifications.

[1] European Parliament created the 95/46/CE Directive (Directive ed9546) with the purpose of homogenising legal cover on data protection, in order to warrant an appropriate protection level on each transfer inside the European Union. At the end of year 2000, the European Parliament extended the personal data regulations initiated by this norm by means of Regulation (CE) 45/2001 (Regulation er45), which covers all that was already established by the Directive 95/46/CE, determines the penalty mechanism at the European level, and creates the figure of the Data Protection European Supervisor as an independent control authority.

In practice, this means that any institution within the European Union context should only store a subset of the personal data, the *relevant data*, that is needed for the purposes of the data collection. The definition of relevance is highly contextual, depending on the activity of the institution and/or the purpose of the archive. Therefore, different institutions will have different definitions of *relevant data*: for instance, relevant data about patients in a hospital clearly should include name, address and any medical information details that are important for the patient's treatment, while relevant data that some companies (e.g. shopping centres) keep about their clients may include name, address and a history of items the client uses to buy (e.g. to adapt stocks and avoid item shortage), but not medical information, as it is not relevant for that company. There are some special cases where data, although being *relevant* for a given institution, is not allowed to be stored. For instance, companies would find useful to have full access to the medical records of their employees, in order to ensure the productivity of its staff by reducing the risk of long-period illnesses. Although in this scenario medical information is relevant, in some European countries that information is not allowed at all or it is only allowed in some specific situations (e.g. with the explicit agreement of the person). In order to ensure personal data protection, all organisations that store and/or process personal data should get a certificate given by a National data protection agency of each EU member state. In such a document there are very specific definitions of which are the *allowed data* and the *allowed data processing* for that particular organisation. Once the allowed data is defined, all regulations on data protection reduce to a single rule: organisations can only store a subset of the relevant data, the *allowed data*, and can only use such data by means of the *allowed data processing*.

Although any EU citizen has the right to access and check the information that any institution has about himself, this is highly impractical. Let us suppose that in the near future any organisation has an agent-mediated eInstitution to provide information and services to individuals and that any person can have an automated personal agent that keeps track of all personal information that organisations have about the person. This agent would enter in each eInstitution checking, for each case, that only *allowed data* is stored, and eventually requesting for amendments or deletions of the information. Such an agent should adapt to the normative and ontological differences between contexts: although the agent may have an ontology defining what *personal data* is, it should be able to adapt its reasoning processes to the regulations and ontologies applying in a given, specific context. For instance, let us focus on two bits of personal information that are protected in the context of European regulations: a person's *blood type* and the person's *race* (Caucasian, Native American, Mongolian, Ethiopian and so on):

- In the generic, *European Union context*, both *blood type* and *race* are *personal data* of a special nature that, in principle, are not *allowed data*, unless some specific regulation or a certificate by a National data protection agency allows the storage and treatment of such information for some specific, well-defined purposes:

  > "Member States shall prohibit the processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life." Article 8.1 in Directive ed9546.

- In the context of any *EU Police Force*, there are special allowances on the use of personal data:

  > "Processing of data relating to offences, criminal convictions or security measures may be carried out only under the control of official authority [...]." Article 8.5 in Directive ed9546.

  That means that, outside the context of an official institution, personal information about criminal antecedents of an individual (a criminal record) is completely forbidden, while in the context of, e.g., the Dutch Police, any relevant information about a criminal or a suspect of a crime (name, address, a physical description -including race- or even medical information such as blood type) is *allowed data*

- In the context of any *EU Health Service*, there are also special allowances on the use of medical data:

  > "Article 8.1 shall not apply where processing of the data is required for the purposes of preventive medicine, medical diagnosis, the provision of care or treatment or the management of health-care services [...]" Article 8.3 in Directive ed9546.

Therefore, inside the context of a hospital information such as *name*, *address* or *blood type* are usually relevant, and belongs to the set of *allowed data* in medical records. On the other hand, *race* is rarely relevant and can only be included in the medical records in those illnesses that are highly related to race.

- The context of the Spanish National Transplants Organisation[2] (ONT) is an interesting, specific subcontext of *EU Health Service*. By Spanish Law, ONT must ensure an equitative and fair distribution of organs and tissues by only taking into account clinical and geographical criteria. Therefore, clinical data such as *blood type* are allowed. Physical descriptions of the donor recipient (basically size, age and weight) are also allowed when they are relevant for the allocation. But such anthropometric data can never include *race*, as it is explicitly forbidden for ONT to use racial information during the allocation process.

In this complex, multi-contextual scenario, a personal agent checking the use of a person's data on each of those contexts does not need to have a full model of all the regulations that apply in a given context. Reasoning on data allowance can be done in an ontological level, that is, the agent should adapt its reasoning to the ontological definitions of *relevant data* and *allowed data* that holds in each context. Some kind of formal model for multi-contextual ontologies is needed, though, in order to properly model the relations between terms in different contexts.

3 Modelling Contextual Ontologies

We will develop our formal framework keeping the following requirements in mind.

1. The formal framework should enable the possibility of expressing lexical differences, because institutions yield terminologies defined on different languages[3]. In particular, in the institutional normative domain, we observe that more concrete contexts mean richer terminologies: talking about personal data comes down to talk about racial data, health data, etc.
2. It should provide a formal semantics (as general as possible) for contextualising terminological expressions.

Following these essential guidelines, a language and a semantics are introduced in this section. The language will make use of part of description logic syntax, as regards the concept constructs, and will make use of a set of operators aimed at capturing the interplay of contexts. In particular, we will introduce:

- A *contextual conjunction* operator. Intuitively, it will yield a composition of contexts: the contexts “personal data in hospitals” and “personal data in police registers” can be intersected on a language talking about data concerning the date of birth and alike generating a common less general context like “anagraphic data in hospitals and police registers”.

- A *contextual disjunction* operator. Intuitively, it will yield a union of contexts: the contexts “personal data in hospitals” and “personal data in police registers” can be unified on a language talking about personal data generating a more general context like “personal data in hospitals or police registers”.

- A *contextual negation* operator. Intuitively, it will yield the context obtained via subtraction of the context negated: the negation of the context “personal data in hospitals” on the language talking about data in general generates a context like “data which are not personal data in hospitals”.

- A *contextual abstraction* operator. Intuitively, it will yield the context consisting in some information extracted from the context to which the abstraction is applied: the abstraction of the context “personal data in hospitals” on the language talking only about anagraphic data generates a context like “anagraphic data in hospitals”. In other words, the operator prunes the information contained in the context “personal data in hospitals” keeping only what is expressible in the language which talks about anagraphic data and abstracting from the rest.

Finally, also *maximum* and *minimum* contexts will be introduced: these will represent the most general, and respectively the least general, contexts on a language. As it appears from this list of examples, operators will need to be indexed with the language where the operation they denote takes place. The point is that con-

[2] The Organización Nacional de Transplantes is a technical organisation within the Spanish Department of Health and Consumer Affairs, whose fundamental mission is the promotion, facilitation and coordination of all types of organs, tissues and bone marrow.

[3] This is a much acknowledged characteristic of contextual reasoning in general, see McCarthy (1986).

texts always belong to a language, and so do operations on them[4].

These intuitions about the semantics of context operators will be clarified and made more rigorous in section 3.2 dedicated to the formal semantics of the framework, and in section 4.1 in which we will formalise an example.

3.1 Language

In a nutshell, the language we are interested in defining can be seen as a meta-language for TBoxes defined on $\mathcal{AL}$ description logic languages, which also handle the union of concepts, full existential quantification (we want to deal with concepts such as “either car or bicycle” and “persons which drive cars”)[5].

The alphabet of the language $\mathcal{L}^{CT}$ (language for contextual terminologies) therefore contains the alphabets of a family of languages $\{\mathcal{L}_i\}_{0\leq i\leq n}$. We take this family to be such that $\{\mathcal{L}_i\}_{0\leq i\leq n} = \mathcal{P}^{+}(\mathcal{L})$, that is to say, each language $\mathcal{L}_i$ is expanded by the “global” language $\mathcal{L}$.

Each $\mathcal{L}_i$ contains a non-empty finite set $\mathbf{A_i}$ of monadic predicates (A), i.e., atomic concepts, and a (possibly empty) set $\mathbf{R_i}$ of dyadic predicates (R), i.e., atomic attributes. These languages contain also concept constructors: each $\mathcal{L}_i$ contains the zeroary operators $\bot$ (bottom concept) and $\top$ (top concept), the unary operator $\neg$ (complement), and the binary operators $\sqcap$ and $\sqcup$. Finally, operators $\forall.$ (universal quantification) and $\exists.$ (existential quantification) apply to attribute-concept pairs.

Besides, the alphabet of $\mathcal{L}^{CT}$ contains a finite set of context identifiers **c**, two families of zeroary operators $\{\bot_i\}_{0\leq i\leq n}$ (minimum contexts) and $\{\top_i\}_{0\leq i\leq n}$ (maximum contexts), two families of unary operators $\{abs_i\}_{0\leq i\leq n}$ (contextual abstraction operator) and $\{\neg_i\}_{0\leq i\leq n}$ (contextual negation operator), two families of binary operators $\{\curlywedge_i\}_{0\leq i\leq n}$ (contexts conjunction operator) and $\{\curlyvee_i\}_{0\leq i\leq n}$ (contextual disjunction operator), one context relation symbol $\preccurlyeq$ (context c_1 “is less general than” context c_2), and finally a contextual subsumption relation symbols “ $. : . \sqsubseteq .$” which is used for both concept contextual subsumption (within context c, concept A_1 is a subconcept of concept A_2) and attribute contextual subsumption (within context c, attribute R_1 is a subattribute of attribute R_2). Lastly, the alphabet of $\mathcal{L}^{CT}$ contains also the sentential connectives $\sim$ (negation) and $\wedge$ (conjunction)[6].

Thus, the set Ξ of context constructs (ξ) is defined through the following BNF:

$$\xi ::= c \mid \bot_i \mid \top_i \mid \neg_i(\xi) \mid abs_i(\xi) \mid \xi_1 \curlywedge_i \xi_2 \mid \xi_1 \curlyvee_i \xi_2.$$

Concept constructs and attribute constructs are defined in the standard way. The set Γ of concept descriptions (γ) is defined through the following BNF:

$$\gamma ::= A \mid \bot \mid \top \mid \neg\gamma \mid \gamma_1 \sqcap \gamma_2 \mid \gamma_1 \sqcup \gamma_2 \mid \forall\rho.\gamma \mid \exists\rho.\gamma.$$

The set P of attributes descriptions (ρ) coincides with the set of all atomic attributes.
The set $\mathcal{A}$ of assertions (α) is then defined through the following BNF:

$$\alpha ::= \xi : \gamma_1 \sqsubseteq \gamma_2 \mid \xi : \rho_1 \sqsubseteq \rho_2 \mid \xi_1 \preccurlyeq \xi_2 \mid \ \sim\alpha$$
$$\mid \alpha_1 \wedge \alpha_2.$$

Technically, a *contextual terminology* in $\mathcal{L}^{CT}$ is a set of subsumption relation expressions on concepts, which are contextualised with respect to the same context. Contextual subsumption relations are the expression by mean of which we give a rigorous characterisation of searlean statements: “X counts as Y in context C”, Searle (1995). This kind of expressions are, in a nutshell, what we are interested in formalising.

In the formalisation of the example, the following symbols will be also used “ $. : . \sqsubset .$” (within context c, concept A_1 is a proper subconcept of concept A_2), and “ $. : . \equiv .$” (within context c, concept A_1 is equivalent to concept A_2). They can be obviously defined as follows:

$$\xi : \gamma_1 \sqsubset \gamma_2 \ =_{def}\ \xi : \gamma_1 \sqsubseteq \gamma_2 \wedge \ \sim \xi : \gamma_2 \sqsubseteq \gamma_1$$
$$\xi : \gamma_1 \equiv \gamma_2 \ =_{def}\ \xi : \gamma_1 \sqsubseteq \gamma_2 \ \wedge\ \xi : \gamma_2 \sqsubseteq \gamma_1.$$

3.2 Semantics

In order to provide a semantics for $\mathcal{L}^{CT}$ languages, we will proceed as follows. First we will define a class of structures which can be used to provide a formal meaning to those languages. We will then characterise the class of operations on contexts that will constitute the semantic counterpart of the context operators symbols introduced in the language. Definitions of the formal meaning of our expressions and of the semantics of assertions will then follow.

[4] Note that indexes might be avoided considering operators interpreted on operations taking place on one selected language, like the largest common language of the languages of the two contexts. However, this would result in a lack of expressivity that we prefer to avoid for the moment.

[5] This type of language is usually referred to as $\mathcal{ALUE}$, or $\mathcal{ALC}$. Within this type of languages the negation of arbitrary concepts is also enabled, see Baader et al. (2002).

[6] It might be worth remarking that language $\mathcal{L}^{CT}$ is, then, an expansion of each $\mathcal{L}_i$ language.

Before pursuing this line, it is necessary to recollect the basic definition of a model for a language $\mathcal{L}_i$, cf. Baader et al. (2002).

Definition 1. (Models for $\mathcal{L}_i$'s)
A model m for a language $\mathcal{L}_i$ is defined as follows:

$$m = \langle \Delta_m, \mathcal{I}_m \rangle$$

where:

- *Δ_m is the (non empty) domain of the model;*
- *$\mathcal{I}_m$ is a function $\mathcal{I}_m : \mathbf{A_i} \cup \mathbf{R_i} \longrightarrow \mathcal{P}(\mathbf{\Delta_m}) \cup \mathcal{P}(\mathbf{\Delta_m} \times \mathbf{\Delta_m})$, such that to every element of $\mathbf{A_i}$ and $\mathbf{R_i}$ an element of $\mathcal{P}(\Delta_m)$ and, respectively, of $\mathcal{P}(\Delta_m \times \Delta_m)$ is associated. This interpretation of atomic concepts and attributes of $\mathcal{L}_i$ on Δ_m is then inductively extended:*

$$\begin{aligned}
\mathcal{I}_m(\top) &= \Delta_m \\
\mathcal{I}_m(\bot) &= \emptyset \\
\mathcal{I}_m(\neg\gamma) &= \Delta_m \setminus \mathcal{I}_m(\gamma) \\
\mathcal{I}_m(\gamma_1 \sqcap \gamma_2) &= \mathcal{I}_m(\gamma_1) \cap \mathcal{I}_m(\gamma_2) \\
\mathcal{I}_m(\gamma_1 \sqcup \gamma_2) &= \mathcal{I}_m(\gamma_1) \cup \mathcal{I}_m(\gamma_2) \\
\mathcal{I}_m(\forall\rho.\gamma) &= \{a \in \Delta_m \mid \forall b, < a, b > \in I_m(\rho) \\
&\quad \Rightarrow b \in I_m(\gamma)\} \\
\mathcal{I}_m(\exists\rho.\gamma) &= \{a \in \Delta_m \mid \exists b, < a, b > \in I_m(\rho) \\
&\quad \& \; b \in I_m(\gamma)\}.
\end{aligned}$$

3.3 Models for $\mathcal{L}^{CT}$

We can now define a notion of *contextual terminology model* (ct-model) for languages $\mathcal{L}^{CT}$.

Definition 2. (ct-models)
A ct-model $\mathbb{M}$ is a structure:

$$\mathbb{M} = \langle \{\mathbf{M_i}\}_{\mathbf{0 \leq i \leq n}}, \mathbb{I} \rangle$$

where:

- *$\{\mathbf{M_i}\}_{\mathbf{0 \leq i \leq n}}$ is the family of the sets of models $\mathbf{M_i}$ of each language $\mathcal{L}_i$. In other words, $\forall m \in \mathbf{M}_i$, m is a basic description logic model of $\mathcal{L}_i$.*
- *$\mathbb{I}$ is a function $\mathbb{I} : \mathbf{c} \longrightarrow \mathcal{P}(\mathbf{M}_0) \cup \ldots \cup \mathcal{P}(\mathbf{M}_n)$. In other words, this function associates to each atomic context in $\mathbf{c}$ a subset of the set of all models in some language $\mathcal{L}_i$: $\mathbb{I}(c) = M$ with $M \subseteq \mathbf{M}_i$ for some i s.t. $0 \leq i \leq n$. Notice that $\mathbb{I}$ fixes, for each context identifier, the language on which the context denoted by the identifier is specified. We could say that it is $\mathbb{I}$ itself which fixes a specific index i for each c.*
- *$\forall m', m'' \in \bigcup_{0 \leq i \leq n} \mathbf{M_i}$, $\Delta_{m'} = \Delta_{m''}$. That is, the domain of all basic description logic models m is unique. We establish this constraint simply because we are interested in modelling different (taxonomical) conceptualisations of a same set of individuals.*

Contexts are therefore formalised as sets of models for the same language, i.e., a set of instantiations of a terminology on that language. This perspective allows for straightforward model theoretical definitions of operations on contexts.

3.4 Operations on contexts

Before getting to this, let us first recall a notion of *domain restriction* ($\rceil$) of a function f w.r.t. a subset C of the domain of f. Intuitively, a domain restriction of a function f is nothing but the function $C \rceil f$ having C as domain and such that for each element of C, f and $C \rceil f$ return the same image. The exact definition is the following one: $C \rceil f(x) = \{y \mid y = f(x) \,\&\, x \in C\}$, cf. Casari (2002).

Definition 3. (Operations on contexts)
Let M' and M'' be sets of models:

$$\begin{aligned}
\rceil_i M' &= \{m \mid m \in M' \\
&\quad \& \; m = \langle \Delta_m, \mathbf{A_i} \rceil \mathcal{I}_\mathbf{m} \rangle\} && (1) \\
M' \Cap_i M'' &= \rceil_i M' \cap \rceil_i M'' && (2) \\
M' \Cup_i M'' &= \rceil_i M' \cup \rceil_i M'' && (3) \\
-_i M' &= \mathbf{M}_i \setminus \rceil_i M''. && (4)
\end{aligned}$$

Intuitively, the operations have the following meaning: operation 1 allows for abstracting the relevant content of a context with respect to a specific language; operations 2 and 3 express basic set-theoretical composition of contexts; finally, operation 4 returns, given a context, the most general of all the remaining contexts. Let us now provide some technical observations. First of all notice that operation $\rceil_i$ yields the empty context when it is applied to a context M', the language of which is not an elementary expansion of $\mathcal{L}_i$. This is indeed very intuitive: the context obtained via abstraction of the context "dinosaurs" on the language of, say, "botanics" should be empty. Empty contexts can be also obtained through the $\Cap_i$ operation. In that case the language is shared, but the two contexts simply do not have any interpretation in common. This happens, for example, when the members of two different football teams talk about their opponents: as a matter of fact, no interpretation of the concept `opponent` can be shared without jeopardising the fairness of the match.

$$abs_0(c_{CONT}) \curlyvee_0 abs_0(c_{PF}) \preccurlyeq c_{SUP} \tag{5}$$

$$c_{SUP} : \texttt{personal\_data} \sqcap \texttt{relevant\_data} \sqsubseteq \texttt{allowed\_data} \tag{6}$$

$$c_{CONT} \curlyvee_1 c_{PF} : \texttt{personal\_data} \sqcap \exists \mathrm{refer}.\texttt{blood\_type} \sqsubset \texttt{relevant\_data} \tag{7}$$

$$c_{CONT} \curlyvee_1 c_{PF} : \texttt{personal\_data} \sqcap \exists \mathrm{refer}.\texttt{anthropometric\_properties} \sqsubset \texttt{relevant\_data} \tag{8}$$

$$\begin{aligned} c_{CONT} \curlyvee_1 c_{PF} : {} & \texttt{personal\_data} \sqcap \exists \mathrm{refer}.\texttt{race} \sqsubset \texttt{personal\_data} \\ & \sqcap \exists \mathrm{refer}.\texttt{anthropometric\_properties} \sqcup \neg \texttt{relevant\_data} \end{aligned} \tag{9}$$

$$c_{CONT} : \texttt{race} \sqsubseteq \neg \texttt{anthropometric\_properties} \tag{10}$$

$$c_{PF} : \texttt{race} \sqsubset \texttt{anthropometric\_properties}. \tag{11}$$

Figure 1: $\mathcal{L}^{CT}$ formalisation of the scenario

3.5 Formal meaning of Ξ and $\mathcal{A}$

The semantics of contexts constructs Ξ can be now defined.

Definition 4. (Semantics of contexts constructs) *The semantics of context constructors is defined as follows:*

$$\begin{aligned}
\mathbb{I}(c) &= M \in \mathcal{P}(\mathbf{M}_0) \cup \ldots \cup \mathcal{P}(\mathbf{M}_n) \\
\mathbb{I}(\bot_i) &= \emptyset \\
\mathbb{I}(\top_i) &= \mathbf{M}_i \\
\mathbb{I}(\xi_1 \curlywedge_i \xi_2) &= \mathbb{I}(\xi_1) \Cap_i \mathbb{I}(\xi_2) \\
\mathbb{I}(\xi_1 \curlyvee_i \xi_2) &= \mathbb{I}(\xi_1) \Cup_i \mathbb{I}(\xi_2) \\
\mathbb{I}(\neg_i(\xi)) &= -_i \mathbb{I}(\xi) \\
\mathbb{I}(abs_i(\xi)) &= \rceil_i \mathbb{I}(\xi).
\end{aligned}$$

As anticipated, atomic contexts are interpreted as sets of models on some language $\mathcal{L}_i$; the $\bot_i$ context is interpreted as the empty context (the same on each language); the $\top_i$ context is interpreted as the greatest, or most general, context on $\mathcal{L}_i$; the binary $\curlywedge_i$-composition of contexts is interpreted as the greatest lower bound of the restriction of the interpretations of the two contexts on $\mathcal{L}_i$; the binary $\curlyvee_i$-composition of contexts is interpreted as the lowest upper bound of the restriction of the interpretations of the two contexts on $\mathcal{L}_i$; context negation is interpreted as the complement with respect to the most general context on that language; finally, the unary abs_i operator is interpreted just as the restriction of the interpretation of its argument to language $\mathcal{L}_i$.

Semantics for the assertions $\mathcal{A}$ and for the contextual concept description $\mathcal{D}$ in $\mathcal{L}^{CT}$ is based on the function $\mathbb{I}$. In what follows we denote with $\delta(\mathcal{I})$ the domain of an interpretation function $\mathcal{I}$.

Definition 5. (Semantics of assertions: $\models$) *The semantics of assertions is defined as follows:*

$$\begin{aligned}
\mathbb{M} \models \xi : \gamma_1 \sqsubseteq \gamma_2 \ & \textit{iff } \forall m \in \mathbb{I}(\xi) : \ \gamma_1, \gamma_2 \in \delta(\mathcal{I}_m) \\
& \textit{and } \mathcal{I}_m(\gamma_1) \subseteq \mathcal{I}_m(\gamma_2) \\
\mathbb{M} \models \xi : \rho_1 \sqsubseteq \rho_2 \ & \textit{iff } \forall m \in \mathbb{I}(\xi) : \ \rho_1, \rho_2 \in \delta(\mathcal{I}_m) \\
& \textit{and } \mathcal{I}_m(\rho_1) \subseteq \mathcal{I}_m(\rho_2) \\
\mathbb{M} \models \xi_1 \preccurlyeq \xi_2 \ & \textit{iff } \mathbb{I}(\xi_1) \subseteq \mathbb{I}(\xi_2) \\
\mathbb{M} \models \sim \alpha \ & \textit{iff not } \mathbb{M} \models \alpha \\
\mathbb{M} \models \alpha_1 \wedge \alpha_2 \ & \textit{iff } \mathbb{M} \models \alpha_1 \textit{ and } \mathbb{M} \models \alpha_2.
\end{aligned}$$

A contextual concept subsumption relation between γ_1 and γ_2 holds iff concepts γ_1 and γ_2 are defined in the models constituting context ξ, i.e., they receive a denotation in those models, and all the basic description logic models constituting that context interpret γ_1 as a subconcept of γ_2. Note that this is precisely the clause for the validity of a subsumption relation in standard description logics, but conditioned to the fact that the concepts involved are actually meaningful in that context. This further condition in the clause is necessary because our contexts have different languages. Perfectly analogous observations hold also for the clause regarding contextual attribute subsumption relations. The $\preccurlyeq$ relation between context constructs is interpreted as a standard subset relation: $\xi_1 \preccurlyeq \xi_2$ means that the context denoted by ξ_1 contains at most all the models that ξ_2 contains, that is to say, ξ_1 is *at most as general as* ξ_2. Note that this relation, being interpreted on the $\subseteq$ relation, is reflexive, antisymmetric and transitive. In Grossi and Dignum (2004) a generality ordering with similar properties was imposed on the set of context identifiers, and analogous properties for a similar relation have been singled out also in Goldman (1976). The interesting thing is that such an ordering is here emergent from the semantics. Note also that this relation holds only between contexts specified on the same language. Clauses for boolean connectives are the obvious ones.

4 Contextual Ontologies at Work

4.1 Formalising an example

We are now able to provide a formalisation of a fragment of the scenario presented in the first part of the paper, making use of the formal semantic machinery just exposed.

Example. (Personal data in transplant organisations and police forces) *We will formalise how the use of personal data is regulated in the two different contexts of Dutch police force (PF) and of the Spanish national transplant organisation (ONT) in accordance with the directives applying to the superordinate European context. We will see how the two concrete contexts PF and ONT implement the same European norm differently: personal data that are allowed to be operated by an institution are only those which are strictly relevant for the execution of the purpose of that institution. The two concrete contexts PF and ONT presuppose a different understanding of what counts as allowed data, because their understanding of the norm lies in divergent ontologies of the concepts involved*[7].

To formalise the scenario a language $\mathcal{L}$ is needed, which contains the following atomic concepts: $\texttt{personal\_data}$, $\texttt{relevant\_data}$, $\texttt{allowed\_data}$, $\texttt{blood\_type}$, $\texttt{race}$, $\texttt{anthropometric\_properties}$*; and the following atomic attribute:* refer. *From this language we obtain* $2^6 - 1 \cdot 2$ *languages* $\mathcal{L}_i$[8]. *Three atomic contexts are at issue here: the context of the superordinate European regulation, let us call it* c_{SUP}*; the contexts of the municipal regulations ONT and PF, let us call them* c_{ONT}, c_{PF} *and* c_{M3} *respectively. These contexts should be interpreted on two relevant languages* $\mathcal{L}_0$*, i.e., the language of the context of European regulation, and* $\mathcal{L}_1$*, i.e., the language of the two concrete contexts PF and ONT. Languages* $\mathcal{L}_0$ *and* $\mathcal{L}_1$ *are such that:*

$$\begin{aligned}
\mathbf{A}_0 &= \{\texttt{personal\_data}, \texttt{relevant\_data}, \\
&\quad \texttt{allowed\_data}\}, \\
\mathbf{R}_0 &= \emptyset
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{A}_1 &= \{\texttt{personal\_data}, \texttt{relevant\_data}, \\
&\quad \texttt{allowed\_data}, \texttt{blood\_type}, \\
&\quad \texttt{race}, \texttt{anthropometric\_properties}\}, \\
\mathbf{R}_1 &= \{\mathrm{refer}\}.
\end{aligned}$$

That is to say, an abstract language concerning only personal, relevant and allowed data, and a more detailed language concerning, besides personal, relevant and allowed data, also blood type, race, anthropometric properties and the refer attribute.

To model the desired situation, our ct-model should then at least satisfy the $\mathcal{L}^{CT}$ *formulas listed in figure 1.*

Formula (5) plays a key role, stating that the two contexts c_{ONT}, c_{PF} are concrete variants of context c_{SUP}. It tells this by saying that the context obtained by joining the two concrete contexts on language $\mathcal{L}_0$ (the language of c_{SUP}) is at most as general as context c_{SUP}. As we will see in the following section, this makes c_{ONT}, c_{PF} inherit what holds in c_{SUP}. Formulas (6)-(11) all express contextual subsumption relations. It is worth stressing that they can all be seen as formalising counts-as statements which specify the ontologies holding in the contexts at issue. Formula (6) formalises the abstract rule to the effect that personal data which are relevant for the accomplishment of the aim of the organisation are allowed to be recorded and used. Formulas (7) and (8) express subsumptions holding in both contexts. Formula (9) tells something interesting, namely that data about race, in order to be used, has to be considered as anthropometric information. Indeed, it might be seen as a clause avoiding “cheating” classifications such as: “data about race counts as data about blood type”. Finally, formulas (10) and (11) describe how precisely the ontologies holding in the two contexts diverge.

4.2 Discussing the formalisation

To discuss in some more depth the proposed formalisation, let us first list some interesting logical consequences of formulas (5)-(11) in figure 2. We will focus on subsumptions contextualised to monadic contexts, that is to say, we will show what the consequences of formulas (5)-(11) are at the level of the

[7] It is instructive to notice, in passing, that no deontics is actually enabled in our formalism. Indeed, the norm according to which only relevant personal data can be operated will be treated as a subsumpion statement. This might be regarded as simplistic, but notice that our attention here does not focus on normative reasoning problems such as reasoning about violations at the level on individuals (ABox), and therefore no deontics is strictly required here.

[8] See section 3.3 in which the language $\mathcal{L}^{ct}$ is presented.

$$
\begin{array}{rcl}
(5),(6) & \models & c_{ONT} : \mathtt{personal\_data} \sqcap \mathtt{relevant\_data} \sqsubseteq \mathtt{allowed\_data} \\
(5),(6),(7) & \models & c_{ONT} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{blood\_type} \sqsubset \mathtt{relevant\_data} \\
(5),(6),(7) & \models & c_{ONT} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{blood\_type} \sqsubset \mathtt{allowed\_data} \\
(5),(6),(8) & \models & c_{ONT} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{anthropometric\_properties} \sqsubset \mathtt{relevant\_data} \\
(5),(6),(8) & \models & c_{ONT} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{anthropometric\_properties} \sqsubset \mathtt{allowed\_data} \\
(8),(10) & \models & c_{ONT} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{race} \sqsubset \mathtt{personal\_data} \\
 & & \sqcap \exists \mathrm{refer}.\neg \mathtt{anthropometric\_properties} \\
(5),(6),(9),(10) & \models & c_{ONT} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{race} \sqsubset \neg \mathtt{relevant\_data} \\
 & & \\
(5),(6) & \models & c_{PF} : \mathtt{personal\_data} \sqcap \mathtt{relevant\_data} \sqsubseteq \mathtt{allowed\_data} \\
(5),(6),(7) & \models & c_{PF} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{blood\_type} \sqsubset \mathtt{relevant\_data} \\
(5),(6),(7) & \models & c_{PF} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{blood\_type} \sqsubset \mathtt{allowed\_data} \\
(5),(6),(8) & \models & c_{PF} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{anthropometric\_properties} \sqsubset \mathtt{relevant\_data} \\
(5),(6),(8) & \models & c_{PF} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{anthropometric\_properties} \sqsubset \mathtt{allowed\_data} \\
(8),(11) & \models & c_{PF} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{race} \sqsubset \mathtt{personal\_data} \\
 & & \sqcap \exists \mathrm{refer}.\mathtt{anthropometric\_properties} \\
(5),(6),(9),(11) & \models & c_{PF} : \mathtt{personal\_data} \sqcap \exists \mathrm{refer}.\mathtt{race} \sqsubset \mathtt{relevant\_data}
\end{array}
$$

Figure 2: Logical consequences of formulas (5)-(11)

two contexts c_{ONT}, c_{PF}. These are indeed the formulas that we would intuitively expect to hold in our scenario. The list displays two sets of formulas grouped on the basis of the context to which they pertain. Let us have a closer look at them; the first consequence of each group results from the generality relation expressed in (5), by means of which, the content of (6) is shown to hold also in the two concrete contexts: in simple words, contexts c_{ONT}, c_{PF} inherit the general rule stating that only relevant personal data can be included and used. Via this inherited rule, and via (7) and (8), it is shown that, in all contexts, data about blood type and anthropometric properties are always allowed. As to data about blood type and anthropometric properties, all contexts agree. Differences arise in relation with how the concept of race is handled. Those differences determine a variation in the interpretation of the abstract norm expressed in (6).

In context c_{ONT}, we have that data about race should not be taken as relevant, and this conclusion is reached restricting the interpretation of what counts as anthropometric information (10) and by means of the "no-cheating" clause (9). In fact, in this context, data about race are not anthropometric data. Context c_{PF}, instead, expresses a different view. Since race counts as anthropometric information (11), data about race are actually relevant data and, as such, can be used.

Before ending the section, we confront this context-based approach with the more standard ones based instead on the defeasible reasoning paradigm. In a non-monotonic reasoning setting, the key point of the example (the fact that the two contexts diverge in the classification of the concept `race`) would be handled by means of a notion of exception: "normally, race is an anthropometric property and is then an allowed type of personal data" and "every exceptional anthropometric property is a forbidden type of personal data". We deem these approaches, despite being effective in capturing the reasoning patterns involved in this type of scenarios, to be inadequate for analysing problems related with the *meaning* of the terms that trigger those reasoning patterns. Those reasoning patterns are defeasible because the meaning of the terms involved is not definite, it is vague, it is -and this is the thesis we hold here- context dependent[9]. Our proposal consists instead in analysing these issues in terms of the notion of context: according to (in the context of) PF race is an anthropometric property; according to (in the context of) ONT race does not count as an anthropometric property. Be-

[9] The issue of the relationship between contextuality and defeasibility has been raised also in Akman and Surav. (1996).

sides enabling the possibility of representing semantic discrepancies, such an approach also has the definite advantage of keeping the intra-contextual reasoning classical, framing non-monotonicity as emergent property at the level of inter-contextual reasoning. Furthermore, the use of description logic allows for its well known interesting computability properties to be enabled at the intra-contextual reasoning level, thus making the framework appealing in this respect as well.

5 Specifying Contextual Ontologies for eInstitutions

In the previous sections we have given an idea of how ontologies and context are used in institutions in order to determine whether or not norms apply to a given situation. We have given a formal framework to formalise the contexts and have shown how this framework can be used to represent and reason about norms in an eInstitution. Although an implementation covering all the aspects of the formal machinery proposed in the previous sections would be computationally expensive, an optimal implementation of the ontological aspects of norms can be far less complex.

It is important to note here that implementing the contextual ontological aspects does not mean implementing some sort of model-checker to verify the formal models of the norms and situations that can be described in a formal framework such as ours, since one is only going to encounter a limited number of contexts at a given time. From the institutional perspective, as we can consider an eInstitution as a single context, all contextual ontological issues are solved during the design process of the eInstitution when defining its ontology. From the agents' perspective, the contextual ontological problems should be solved on-line; agents that are joining the eInstitution need to know in which context they are supposed to work, and need to be informed of the ontology and norms applicable in the eInstitution.

From the eInstitution's point of view, the ontological aspects of norms mainly impact two steps in the eInstitution's implementation: a) the definition of the *eInstitution's ontology*, giving an interpretation of all the terms in the norms, and b) the implementation of the *norm enforcement mechanisms*, following the norm interpretation given by the ontology.[10] From the ontological perspective, the most complex step is the definition of its ontology, as several contextual ontologies should be taken into account. That is, not only does one need to look at the concepts and norms necessary for the eInstitution's context, but one also has to consider the (super)contexts in which the eInstitution is to operate, which are possibly nested (e.g., regional/national/transnational/international contexts). In practice, this means that one needs to create some kind of link from the ontologies of different supercontexts to the institutional ontology. In our approach (which is ongoing work), the links between ontologies are explicitly defined by the designer by means of different kinds of ontology abstraction and ontology inheritance relations. The simplest scenario is when an eInstitution has a set of non-conflicting nested supercontexts. For instance, in the case of an eInstitution for the Spanish National Transplant Organisation (ONT), in order to define ONT's ontology we can inherit terms from its supercontexts: The Spanish National Health System, the Spanish Law and the European Union Law. It is important to note that an explicit link for all inherited terms should be kept in the ontologies' representation. Then the inherited terms can be extended in ONT's ontology with extra terms and/or re-defined, if needed, for the particular context of the institution. A more complex scenario appears when an eInstitution has disjoint nested supercontexts with conflicting definitions of terms. This is the case of transnational institutions such as Eurotransplant[11], where different ontological definitions of terms may appear in each of the countries where the institution should operate. In this case, when inheriting different, conflicting definitions of the same term into the ontology, the designers should solve the conflict by precisely agreeing on and defining the precise meaning of the term that will apply inside the context of the eInstitution.

From the individual agents' perspective, the ontological aspects of norms and the issue of multi-contextual ontologies influences the on-line reasoning cycle of the agent. That is, when an agent tries to enter an eInstitution it is told which ontologies and norms are used in the eInstitution. However, the ontology used by the eInstitution need not be the same as that of the agent, and concepts in the norms used in the eInstitution might be unclear to the agent. In this case, the eInstitution and agent need to obtain a common understanding of the concepts such that it provides the agent with a clear meaning of the norms used in the institution. This can be done by finding

[10] More details on the implementation of norm enforcement mechanisms can be found in Vázquez-Salceda et al. (2004).

[11] The Eurotransplant International Foundation is responsible for the mediation and allocation of organ donation procedures in Austria, Belgium, Germany, Luxembourg, the Netherlands and Slovenia.

a common supercontext and using the ontology's abstraction and inheritance relations to this supercontext.

6 Conclusions and Future Work

The motivating question of our research was how institutions make their norms operative in the domain they are supposed to regulate, i.e., how do institutions implement norms. The thesis we held here is that institutions are based on ontologies. Via these ontologies they translate norms, which are usually formulated in abstract terms (for instance, the concept of "relevant data"), into concrete constraints which are instead understandable in the terms used to describe the situations they regulate (for example, "data about blood type"). As institutions are supposed to regulate completely different domains, the ontologies they are based on are also different. They can be specified on completely different vocabularies, or, if they share a set of terms, they may interpret it in divergent ways (which is the case of the concept of "relevant data" we discussed in our example). To get a grip on this phenomenon, we made use of contexts as means to localise these ontological discrepancies: institutions are based on ontologies, and these ontologies are contextual. This is also the analytical setting in which we provided a clear understanding of the so called *counts-as* phenomenon; counts-as statements are nothing but contextual subsumption relations: they are the basic brick by means of which institutions establish their ontologies.

This analysis has then been framed in a rigorous setting. The formal framework exposed is based on a specific understanding of the notion of *context* as set of models for particular description logic languages, and provides a formal characterisation of the notion of *contextual ontology*. This framework is also used for formalising an example. At the end of the paper we also provided some general ideas on how these contextual ontologies can be concretely used in order to specify and reason about eInstitutions.

With respect to future work, we firstly intend to develop an extension of the framework which can enable a full-fletched interaction of context and ontological features with the more standard normative reasoning issues (eminently, reasoning with violations). This requires to focus also on aspects concerning reasoning with instances of concepts (what in description logics is called ABox), and of course on the inclusion of some deontic logic. Secondly, this extension should be brought into practice and applied in the development of eInstitutions and Normative Agents.

References

V. Akman and M. Surav. Steps toward formalizing context. *AI Magazine*, 17(3):55–72, 1996.

C. E. Alchourrón and E. Bulygin. *Normative Systems*. Springer Verlag, Wien, 1986.

F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider. *The Description Logic Handbook*. Cambridge University Press, Cambridge, 2002.

G. Boella and L. Van der Torre. Regulative and constitutive norms in normative multiagent systems. In *Proceedings of KR2004, Whistler, Canada*, pages 255–266, 2004.

J. Broersen, F. Dignum, V. Dignum, and J.-J. Ch. Meyer. Designing a Deontic Logic of Deadlines. In *7th Int. Workshop on Deontic Logic in Computer Science (DEON'04)*, Portugal, 2004.

E. Bulygin. On norms of competence. *Law and Philosophy 11*, pages 201–216, 1992.

E. Casari. *La Matematica della Verità*. Privately distributed, 2002.

F. Dignum. Abstract norms and electronic institutions. In *Proceedings of the International Workshop on Regulated Agent-Based Social Systems: Theories and Applications (RASTA '02), Bologna*, pages 93–104, 2002.

F. Dignum, J. Broersen, V. Dignum, and J.-J. Ch. Meyer. Meeting the Deadline: Why, When and How. In *3rd Goddard Workshop on Formal Approaches to Agent-Based Systems (FAABS)*, Maryland, 2004.

F. Dignum, D. Kinny, and L. Sonenberg. From Desires, Obligations and Norms to Goals. *Cognitive Science Quarterly*, 2(3-4):407–430, 2002.

V. Dignum, J.-J.Ch. Meyer, F. Dignum, and H. Weigand. Formal Specification of Interaction in Agent Societies. In *2nd Goddard Workshop on Formal Approaches to Agent-Based Systems (FAABS)*, Maryland, Oct. 2002.

Directive ed9546. Directive 95/46/CE of the European Parliament and of the Council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and of the free movement of such data, October 1995.

Eurotransplant. Eurotransplant International Foundation.
http://www.eurotransplant.nl.

F. Giunchiglia and P. Bouquet. Introduction to contextual reasoning. *Perspectives on Cognitive Science*, 3, 1997.

A. I. Goldman. *A Theory of Human Action*. Princeton University Press, Princeton, 1976.

D. Grossi and F. Dignum. From abstract to concrete norms in agent institutions. In *Proceedings of FAABS III Workshop, Washington, april*, 2004.

D. Grossi, F. Dignum, and J-J. Ch. Meyer. Contextual taxonomies. In J. Leite and P. Toroni, editors, *Proceedings of CLIMA V Workshop, Lisbon, September*, pages 2–17, 2004a.

D. Grossi, F. Dignum, and J-J. Ch. Meyer. Contextual terminologies. Draft, 2004b.

H. L. A. Hart. *The Concept of Law*. Clarendon Press, Oxford, 1961.

A. J. I. Jones and M. Sergot. Deontic logic in the representation of law: towards a methodology. *Artificial Intelligence and Law 1*, 1992.

A. J. I. Jones and M. Sergot. On the characterization of law and computer systems. *Deontic Logic in Computer Science*, pages 275–307, 1993.

J. McCarthy. Notes on formalizing contexts. In Tom Kehler and Stan Rosenschein, editors, *Proceedings of the Fifth National Conference on Artificial Intelligence*, pages 555–560, Los Altos, California, 1986. Morgan Kaufmann.

J. McCarthy. Generality in artificial intelligence. *Communications of the ACM*, 30(12):1030–1035, 1987.

A. Peczenik. *On Law and Reason*. Kluwer, Dordrecht, 1989.

Regulation er45. Regulation (EC) No. 45/2001 of the European Parliament and of the Council of 18 december 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data. Official Journal of the European Communities, January 12 2001.

A. Ross. *Directives and Norms*. Routledge & Kegan Paul, London, 1968.

J. Searle. *The Construction of Social Reality*. Free Press, 1995.

J. Vázquez-Salceda. *The Role of Norms and Electronic Institutions in Multi-Agent Systems. The HARMONIA framework*. Whitestein Series in Software Agent Technology. Birkhäuser Verlag, 2004.

J. Vázquez-Salceda, H. Aldewereld, and F. Dignum. Implementing norms in multiagent systems. In G. Lindemann, J. Denzinger, I.J. Timm, and R. Unland, editors, *Multagent System Technologies*, LNAI 3187, pages 313–327. Springer-Verlag, 2004.

G. H. Von Wright. *Norm and Action. A Logical Inquiry*. Routledge, London, 1963.

Self-Organized Criticality and Norm Avalanches

Matthew J. Hoffmann
Department of Political Science and International Relations
University of Delaware
Newark, DE 19716, USA
mjhoff@udel.edu

Abstract

Norm emergence and evolution remain crucial, open questions in international relations. This paper suggests that a self-organized criticality perspective may prove fruitful in the pursuit of understanding social norm dynamics. The paper presents an agent-based model that formalizes the norm life cycle proposed by Martha Finnemore and Kathryn Sikkink (1998), and simulates norm emergence and evolution. The results of simulation exercises demonstrate that the norm life cycle is a plausible mechanism for explaining norm emergence and evolution, and they reveal potential conditions under which norms emerge and evolve. Further, analysis of the results suggests that the simulated norm 'avalanches' follow the power law distributions expected in self-organized critical system.

1. Introduction

In this essay I offer an agent based model called "Pick a Number" that captures both norm emergence and evolution. The model formalizes a specific social constructivist framework from the international relations literature—Finnemore and Sikkink's (1998) norm life cycle—and explores the role that norm entrepreneurs play in catalyzing the emergence and evolution of social norms. The results of simulation exercises demonstrate that the norm life cycle is a plausible mechanism for explaining norm dynamics in world politics and perhaps beyond. Further, initial analyses of the results also suggest that the simulated norm 'avalanches' follow the power law distributions expected in self-organized critical systems. The paper thus provides a foundation for a self-organized criticality perspective with which to explore social norms theoretically and empirically.

2. Social Norms

Norm has become a ubiquitous term in the lexicon of international relations, political science, and the social sciences more generally (see e.g. Ensminger and Knight, 1997; Ostrom, 2000; Epstein, 2000). However, explaining the emergence and evolution of norms has been a more arduous task. One particular framework—Finnemore and Sikkink's (1998) norm life cycle—appears to have a great deal of potential, especially for exploring international social norms, but has yet to be formally explored.

As constructivists, Finnemore and Sikkink begin by defining social norms as standards of "appropriate behavior for actors with a given identity." (1998, 891) They posit that agents base their behavior on the logic of appropriateness—determining a course of action based upon what is appropriate given their identity (March and Olsen 1989). Because norms play a central, causal role in constructivism, a good deal of work has surrounded the emergence and evolution of norms. Finnemore and Sikkink's proposed norm life cycle explores these dynamics through three linked stages: emergence, cascade, and internalization (1998, 896-901).

Norm emergence begins with a catalyst—a norm entrepreneur. Norm entrepreneurs are agents that, dissatisfied with their social context, advocate different ideas about appropriate behavior from organizational platforms that give their ideas credence. These norm entrepreneurs work to persuade other agents to alter their behavior in accordance with the entrepreneur's ideas of appropriateness. For constructivists, norm entrepreneurs attempt to alter what agents think is appropriate behavior. How this alteration takes place is currently a matter for debate in the constructivist literature, but when a 'critical mass' of agents has accepted the new ideas as appropriate or a threshold of norm acceptance is passed, then Finnemore and Sikkink claim that a norm has emerged (1998, 901). Following emergence, the nascent norm cascades throughout the system (1998, 902). The final stage in the cycle is internalization. The norm becomes taken for granted, and conformance with its dictates is no longer (or rarely) questioned (Finnemore and Sikkink 1998, 904; See also Epstein 2000).

Implicit in this framework are the defining characteristics of a driven threshold system (see, e.g. Cederman, 1997; Bak and Chen, 1991). Indeed, Finnemore and Sikkink are describing the dynamics of self-organized criticality (see e.g. Bak and Chen, 1991, Cederman, 1997, 2003; Brunk 2002). The suggestions of the norm entrepreneur are the dropping sand grains from the canonical metaphor of self-organized criticality—the sand-pile model. A norm entrepreneur's suggestions represent a stimulus to a population of interdependent, adaptive agents and this stimulus potentially catalyzes a number of different responses: 1.No change—the agents are resistant to the stimulus, continue their individual behavior and the current social patterns and relationships continue unchanged. 2. Limited cascade—some agents change their individual behavior, but their altered behavior is insufficient to significantly alter the current social patterns and relationships. 3. Substantial cascade—some (or many) agents change their individual behavior, and their altered behavior transforms the current social patterns and relationships. Substantial cascades could signal normative breakdown (if the current social pattern is structured by a stable norm) or normative emergence (if the current social pattern is norm contestation or lack of a norm). The size of the norm cascade—the spread of a norm to the whole population and the length of time a norm remains in force—and the stability of existing norms is dependent upon the connections and interconnections of the individual agents as well as the characteristics of the stimulus itself (how clearly communicated, normative value, power of the entrepreneur). The task of the model is to explore if and when entrepreneurs can catalyze a cascade of norm emergence or change.

3. The Model—Pick a Number

This study is based on an approach to modeling that begins with verbal models of political phenomena. I formalized the norm life cycle in order to rigorously explore its deductions and conclusions, generate new hypotheses, and build theory. Thus the model described below does not explicitly draw on existing computational models of social norms (e.g. Conte and Castelfranchi, 1995; Saam and Harrer, 1999), but instead brings agent-based modeling to a verbal framework that has yet be formalized. The model (written in C++ using Microsoft Visual Studio—the code is available on request) explores the role that a norm entrepreneur can play in catalyzing norms and how norms can ebb and flow in a self-organized critical system. The model works with a verbal framework that is explicitly drawn from the international relations literature and is thus potentially constrained by the understanding of norms in international relations (a point I will return to below).

The model results demonstrate that under certain conditions, norm entrepreneurs can in fact catalyze norm emergence *and* norm change, lending credence to the constructivist claims and further that a self-organized criticality perspective provides significant insights into norm dynamics.

The model simulates 10 agents who each pick a number between 0 and 100 in an attempt to match the group outcome, which I have defined to be the average (arithmetic mean) of the choices from the entire group. This foundation is designed to capture the logic of appropriateness and the mutual constitution of agents and structures from constructivist theorizing. Agents in constructivist theory and in this model strive to behave appropriately, which is defined by the social context (group outcome), which in turn is constructed from the behaviors of the agents within it (aggregated choices of individual agents).

The tools available to the agents for making their predictions or picking their behaviors are (very) simple rules. In the simulations presented the agents have available to them a universe of seven rules. The rules themselves are mutually exclusive random choices drawn from a uniform distribution of integers within specified boundaries: Rule 1: 0 – 10, Rule 2: 15 – 25, Rule 3: 30 – 40, Rule 4: 45 – 55, Rule 5: 60 –70, Rule 6: 75 – 85, Rule 7: 90 – 100. Each agent is initially randomly assigned three of these prediction rules (without repeats). An agent uses one of these three rules—the public rule—to make the prediction that is sent to the entire group. Each agent determines which rule is public by keeping track of scores for each rule in its repertoire. Each rule starts with a baseline score of 100 and the score rises and falls depending on how close its predictions have been to the group outcome. The rule with the currently highest score is the public rule.

In order to judge their satisfaction with their rules the agents evaluate the behavior produced by the public rule as well as the potential behavior of their other two private rules. Once the group outcome is known, agents compare their three predictions (one public, two private) with the outcome and reward or penalize their rules depending on the closeness of the prediction. In this model 'close enough' is governed by a parameter called precision. In most runs of the model, precision is set at 5%. This means that rules that predict the group outcome within +/- 5 are rewarded (+1) and others are punished (-1). A private rule becomes public when its score exceeds that of the current public rule. In order to facilitate adaptation and change over time, at set intervals (10-20 rounds) each agent discards a poorly performing rule and is randomly assigned a new rule from the universe of rules. The new rule starts with a fresh score of 100.

The agents' social context is very limited and agents only perceive the group outcome. This characteristic is designed to mimic the limited sociality of agents in world politics. However, it can also be considered that the model does not explicitly simulate how agents obtain an understanding of their social environment—communication and other social activities—and instead focuses exclusively on what happens once agents have a picture of their social environment. This leaves the focus squarely on how a norm entrepreneur can catalyze norm emergence and change.

The catch for the agents in this simple world is that it is a noisy world. While the true outcome is exactly the average of the predictions from the population, the outcome that each agent perceives is obscured by noise (a random draw from a uniform distribution bounded by zero and the specified maximum noise level). The noise can be thought of in two ways. First, it could be simulating a lack of information or uncertainty. Second, it could be conceived as representing the complexity of the social environment—the higher the noise levels, the less clear agents are on what the appropriate group outcome should be.

An additional aspect of the social context is the existence of a natural attractor in this system. Rule 4, which produces predictions between 45 and 55, is a pre-ordained focal point. Averaging random numbers between 0 and 100 will produce a mean of around 50 in the long run—and thus agents should be drawn to this rule. The baseline model explores the conditions under which the agents can find this natural attractor through uncoordinated, adaptive behavior. From there, the real test of the constructivist framework begins and norm entrepreneurs are introduced into the model. Norm entrepreneurs suggest a rule to the agents at specified intervals (every 50 rounds). The model thus becomes a driven threshold model as the norm entrepreneurs periodically introduce suggestions to the agents. Each agent replaces its currently worst performing rule with the norm entrepreneur's suggestion, and the suggested rule starts with a fresh score of 100.

In the base version of the model the entrepreneur is able to reach all agents simultaneously and automatically convinces all the agents in the simulation to add the suggestion to their repertoire of rules (in the sensitivity analysis, (see Hoffmann, forthcoming) I let the reach of the norm entrepreneur vary). Crucially, the agents will only use the suggested rule if their other rules have been weakened through past punishments—i.e. just because a new idea about appropriate behavior is presented, that does not mean it will automatically influence behavior. This model thus tests the importance of norm entrepreneurs for catalyzing norms, though it glosses over issues of how a norm entrepreneur convinces each particular agent.

3.1 Results I – Norm Emergence and Change

When norm entrepreneurs are absent from the system, two types of macro patterns emerge in the simulations. Depending upon the noise levels in the system, the simulation exhibits a strict dichotomy between stability and volatility in the system. When the noise level is low enough (<9%), the agents eventually hit upon the dominant rule in the system, rule 4. As the noise increases (>9%), the agents are unable to come to agreement on any rule and the average prediction reflects this uncertainty. These results, robust to a number of sensitivity analyses, show that the agents' actions produce either a stable (Figure 1) or volatile (Figure 2) macro-pattern with a strict breakpoint between the two types of patterns.[1] The macro patterns, in turn, alter/reinforce agent behavior leading to cycling in rule use or the domination of a single rule. Both of these patterns have significant analogues in politics, and the model suggests that there are conditions (low noise) when norm entrepreneurs are entirely unnecessary for normative emergence and there may be conditions (high noise) when norm entrepreneurs will not be able to catalyze norm emergence.

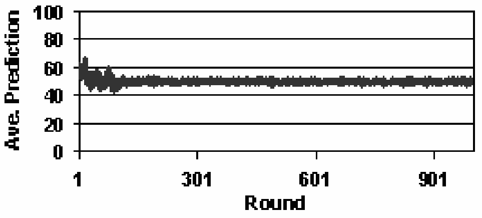


Figure 1: Stability with Low Noise

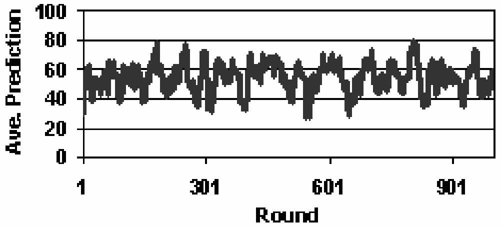


Figure 2: Volatility with High Noise

[1] The figures report the average prediction (group outcome) for each round of the simulation.

In contrast to the dichotomous patterns exhibited when the system lacks norm entrepreneurs, their presence creates different patterns. First, norm entrepreneurs are able to influence which rule rises to dominant status when the noise/precision levels would otherwise lead to stability around the dominant rule (Figure 3). At low levels of noise, the addition of a norm entrepreneur makes it possible for the agents to crystallize around any of the 7 rules. This pattern, too, has important analogues in politics and the model suggests how norm entrepreneurs can foster lock-in around a rule.

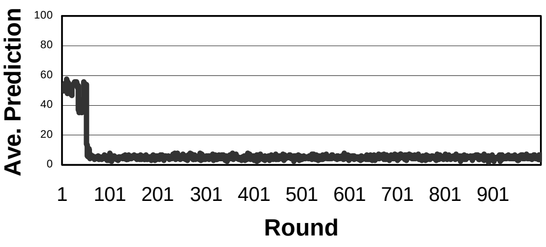


Figure 3: Stability with a Norm Entrepreneur

Second, and more importantly, norm entrepreneurs lead the system into metastable patterns where neither volatility nor stability reigns (Bak and Chen, 1991). Metastable patterns occur when pockets of stability arise but do not last—there is stability in the system but it is not robust. In these simulations, the agents can coalesce around any of the rules and we see the rise and demise of intersubjective agreement among the agents. The norm lasts for a while before eroding via agent choices and new norm entrepreneur suggestions. The stability erodes because the system is too noisy to support long-term stability and norm entrepreneurs periodically prod the system with new suggestions. Figure 4 demonstrates the impact of norm entrepreneurs on a simulation with relatively high noise (10%). Figure 5 displays the rule usage in the entire population, for the same simulation, displaying the percent of the agent population using each rule (publicly) in every round.

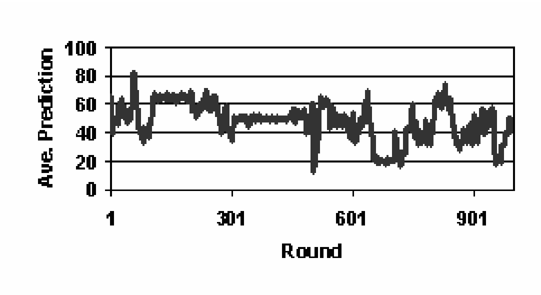


Figure 4: Metastability with Norm Entrepreneurs

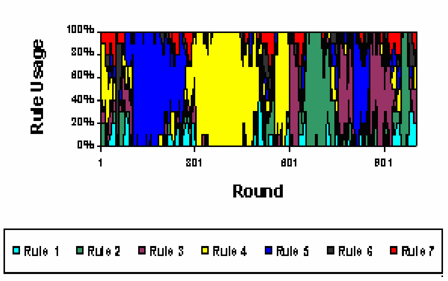


Figure 5: Rule Usage in the Population

These figures demonstrate that metastable patterns result from the norm entrepreneur's suggestions at a level of noise high enough to normally cause volatile outcomes. The norm entrepreneurs catalyze periods of intersubjective agreement among the agents—they make it possible for agents to crystallize around a rule for relatively short periods in a noisy environment. Norm entrepreneurs cause the system to walk a metastable line between volatility and stability, creating the patterns of cascading norms over time that constructivists recognize empirically in their studies of norm emergence and change. The model thus demonstrates that the constructivist expectations for norm entrepreneurs are indeed plausible. The constructivist norm life cycle does produce a system of emerging and evolving norms. Delineating in detail the specific boundary conditions for norm emergence and evolution as well as their empirical analogues is beyond the scope of this paper, but has been explored (Hoffmann, forthcoming).

3.2 Results II – Self-Organized Criticality

Beyond demonstrating the plausibility of the norm life cycle, the modeling results also support the notion that the norm life cycle exhibits self-organized criticality. Indeed, while the norm life cycle verbal model appears to be built with self-organized criticality (SOC) insights, the qualitative results alone do not demonstrate the utility of a SOC perspective for studying social norms. The analysis does provide reasons to *suspect* that social norms are SOC phenomena:

1. The norm life cycle is implicitly a driven threshold model.
2. Phase transitions are evident in the qualitative results.
3. A metastable phase emerged with the introduction of a norm entrepreneur.

However, testing the claim that the norm life cycle is a SOC framework and further that studying norms empirically would benefit from a SOC perspective is more difficult. There is no empirical target available for a power law analysis[2] of social norms that would provide evidence for SOC dynamics, and while empirical testing is not impossible, quantifying social norms in a population is far from trivial and such data is not now available. On the other hand, the modeling exercises readily produce data for power law analysis, yet it is not obvious what should be tested. What aspects of social norms should we expect to be governed by power laws? Multiple aspects of norms are potential candidates for testing—distance from the natural norm, time in between norms, and rate of successful entrepreneurship, to name a few. In this analysis, I focus on the length of time individual norms are in force, an empirically relevant concept because of observed variation in norm lengths (fleeting fads and long-lived moral 'laws').

Focusing on norm lengths requires a clear definition for when a norm is in force and clear decisions on how to produce the norm length data. For this analysis I count a norm being in force when 70% of the population is following the same rule. Data set choice is more difficult as there are an infinite number of potential configurations for producing data and decisions on how many runs to include and the number of rounds that constitute a run are essentially arbitrary. This analysis reports on four data sets generated by the Pick a Number model:

- Data Sets 1-3 (Figures 6-8) contain the results of simulations with norm entrepreneurs present, a single level of social complexity and multiple runs. Each data set contains the norms that emerged in 5 runs (differing in random number seeds) of the model lasting 30,000 rounds per run. Each set is characterized by a different level of social complexity. Such a configuration simulates norm emergence in a single-issue area (specific level of social complexity) across multiple populations (5 different runs). Set 1 explores five runs of the model at low levels of social complexity. In sets 2 and 3 the social complexity increases to medium and high respectively.
- Data set 4 (Figure 9) combines the results of the simulations run for sets 1-3 in a master set. The set collects the norms that emerge in multiple issue areas (varying levels of social complexity) and multiple populations (multiple runs of the model). This master set gives an overall picture of the NLC dynamics across a spectrum of issues and population.

Absolute frequency analysis was performed on each data set to determine the existence of power laws—regressing (linear) the log of the length of time individual norms were in effect and the log of the frequency of norms with a particular length of time in force. If a power law is evident in the data, the linear model should fit the data quite closely. Table one summarizes the analysis.

In three of the four data sets, the intercept (A) and slope (B) coefficients obtained high t-ratios (in parentheses) and highly significant F statistics (p<.0001), suggesting correspondence with a power law model. Only in data set number one, do the results point to a lack of a governing power law. The data sets drawn from simulations with medium and high social complexity (2 and 3), as well the data set that incorporates varied levels of social complexity (4) all exhibit power law behavior. These results demonstrate that the norm life cycle, a candidate explanation for norm dynamics, can also be shown to be a model of SOC. Such a finding suggests that a complex systems perspective on social norms may prove fruitful theoretically and empirically. The results provide a firm foundation for a new research agenda on social norms informed by SOC insights, and they suggest that social norms (empirically) may be governed by power laws.

Table 1: Regression Results

Data Set	F	A	B	R^2
1	6.6 (.015)	.313 (4.7)	-.087 (-2.6)	.149
2	410 (.000)	2.573 (25.1)	-1.032 (-20.3)	.708
3	513 (.000)	3.62 (28.8)	-1.689 (-22.7)	.835
4	270 (.000)	2.173 (20.4)	-.946 (-16.4)	.589

[2] On the use of power law analysis to explore self-organized critical systems see Bak and Chen (1991) and Cioffi-Revilla (in preparation).

However, while the statistical results are evidence that a power law governs the simulated norms in the Pick a Number model, visual inspection uncovers some significant deviations from an ideal power law model in all of the data sets. In data set one, this is far from surprising as the statistics themselves are weak at best. The log-log graph of data set one (Figure 6) immediately rules out a power law distribution given the lack linearity.[3] At low levels of social complexity the model produces a few norms of enormous length (one per run) and very few of any other norms.

While the statistical analysis for the other three data sets is more promising, some bending off of an ideal power law graph is evident in the graphs for all three data sets. The bending is perhaps most severe in Figure 8, where the social complexity is the highest. This data set also produced the most constricted scope—with the observations barely spanning 2 orders of magnitude. In this case, the noise or social complexity in the system is too high to support enough norms of any length (bending at the lower values) and long norms are almost non-existent(the maximum norm length was only 257, in comparison with 4738 for data set 2 and 29,949 for data set 4). In the remaining data sets, the spread of the observations is more expansive, upwards of 3 orders of magnitude, but bending is still apparent. These graphs (Figures 7 & 9) have significantly less bending at the lower values, but still bend in the extreme values.

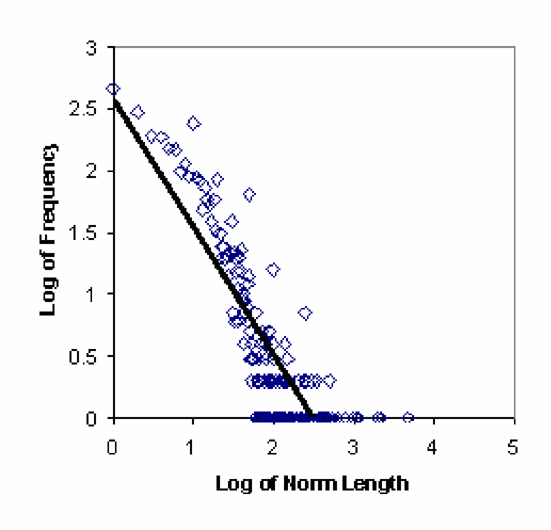


Figure 6: Low Social Complexity

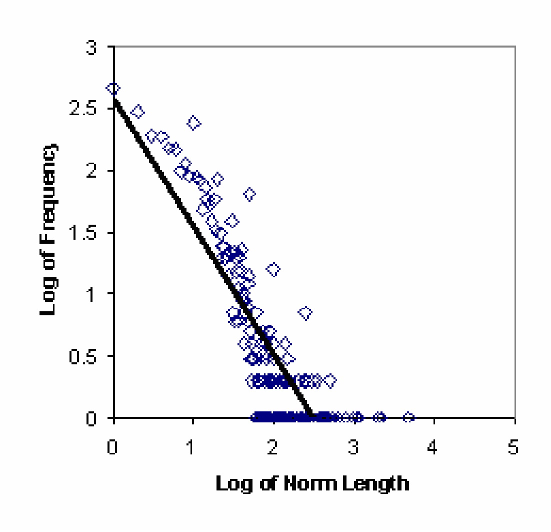


Figure 7: Medium Social Complexity

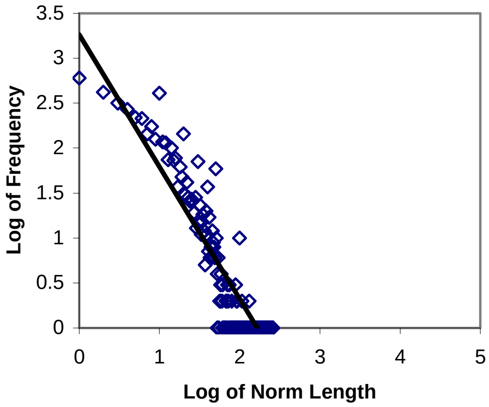


Figure 8: High Social Complexity

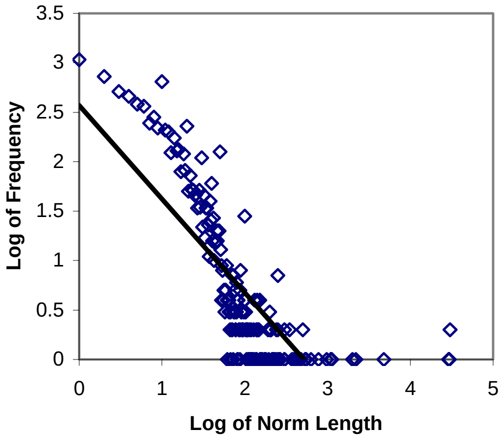


Figure 9: Combined Data Sets

[3] For each of these log-log graphs, the log of the frequency is on the Y-axis and the log of the norm length is on the X-axis. The squares thus represent how often we see a norm of a particular length.

It is somewhat difficult to pinpoint the source of the deviations. Certainly social complexity plays a large role. Examining the three phases individually (data sets 1-3), it is clear that the metastable phase best matches the power law model. Here we have visual linearity (with less bending), a nice spread of norm lengths, and solid statistics. Notice also that in this phase, the power law is close to the classic Zipf distribution (B=-01.032).[4] Repeated trials in this phase show that B hovers around –1.[5] The other phases do not correspond nearly as well, suggesting that social norms in particular issues may be governed by power laws under certain conditions. However, when all three levels of social complexity are included—better approximating a social system with multiple issues—the combined data with all three phases evident conforms to a power law distribution.

Given the artificial nature of the data, the other source for deviation is obviously the model itself—especially the length of the runs. I arbitrarily set the number of rounds in a run to 30,000 and the number of runs at 5. This has very little influence at high or low levels of complexity. In both cases, the length of the run is immaterial. Longer runs at high complexity is unlikely to generate the longer norms that would cure the bending, and at low complexity levels, longer runs would merely lengthen the few locked-in norms. However, in the metastable phase (Data set 2), it is likely that such a limit affects the number of long norms that appear in the data set. Increasing the number of runs (more than 5) and increasing the length of the runs (past 30,000) may produce a cleaner power law graph for the metastable phase.

In the collapsed data (Figure 9) that takes into account all levels of social complexity and includes more runs, we still see bending, but also the possibility of a more ideal power law model. This data set, in that it contains the long norms from the low social complexity runs and the fleeting norms of the high social complexity runs, has the potential to most closely match the power law model. In fact, when this data is grouped by norm length into 8 categories (log length of: (0-1.25), (1.25-1.75), (1.75-2.25), (2.25-2.75), (2.75-3.25), (3.25-3.75), (3.75-4.25), (4.25-4.75)), the bending disappears—see Figure 10.

While deviations are apparent, there appears to be enough evidence to conclude (at least preliminarily) that the norm life cycle produces norms governed by a power law model. The implications of this finding are, unfortunately, not immediately apparent. I began with a verbal abstraction and created a computational abstraction from it. I then gathered and analyzed data from this second order abstraction. What can be learned from this?

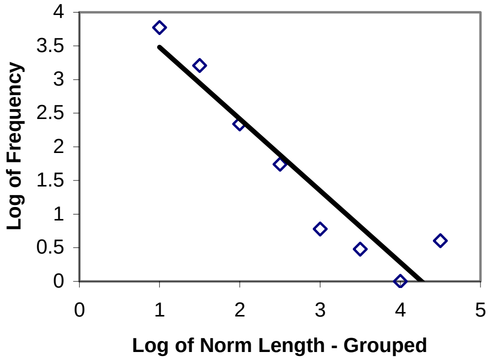


Figure 10: Collapsed Data

First, it is interesting to consider the contrast of this analysis with most power law research. In the Pick a Number model, the avalanches are cascades of stability—when a norm is in force, all of the agents are content, so to speak, behaving appropriately, having their expectations met, predicting the 'correct' number. In many applications (earthquakes, wars, sand piles, forest fires, social protests) the avalanches are of instability. Tension is built up through repeated input into the system and released in power law governed cascades. In the norm life cycle, tension also builds—the agents get more and more frustrated as they cannot predict the correct number (act appropriately)—but the entrepreneur's input leads to a release that is actually the emergence of stability and a self-reinforcing equilibrium. Generalizing beyond norms, therefore, a SOC perspective may provide valuable insight into the emergence of social order in multiple guises—institutions, organizations, economies, and polities.

This analysis provides a very clear assessment of the logic of the norm life cycle. It is clear that the framework entails norm emergence and evolution, as posited by Finnemore and Sikkink. Further, the it entails SOC; it is a driven threshold model. As the norm life cycle generates norms that are governed by power laws, it is at least plausible that the framework's original target (empirical norms) is so governed as well. Thus this analysis provides impetus for empirical research agenda on social norms that is informed by the insights of SOC.

[4] For more on Zipf distributions, see Cioffi-Revilla (in preparation).

[5] Analyses that focus specifically on the metastable region—single runs of extended length (100,000 rounds)—also found similar results.

4. Extensions and Future Research

4.1 Empirical Work

Modeling alone will not advance our knowledge of norm dynamics. What is necessary is a recursive, methodologically eclectic approach. This modeling exercise began with a verbal model and moved to a formal computer model. The next step is empirical testing of the model insights. This research has begun with the initial model results (see Hoffmann, forthcoming), but more is necessary. Especially relevant is empirical examination of the power law findings. Empirical research is not the end either, however. The empirical results should be used to enhance our models. Using both formal and empirical methods is the best way of moving our knowledge forward.

4.2 Norm Competition

As one example, this recursive approach has produced an area for further modeling work. I used the results of the initial model analysis (Results I) to provide insight for a study of participation in the global climate change regime. In so doing, it was clear that while the norm life cycle provides a solid explanation of norm dynamics, it does not capture the whole story of norm dynamics. What is missing from the norm life cycle (and apparent empirically) is norm competition.

The empirical work with climate change, leads me to consider three types of norm competition that should be modeled for further exploration. First, no single norm can be totalizing (as they are represented in the single norm model, Pick a Number) because a variety of norms may be relevant to any given situation. There thus may be a contest among distinct norms. The resulting impact on intersubjective reality is thus intricate and complex. Hybrid norms may appear. Norm complexes may be built. Individual norms may rise and fall through direct confrontation.[6] Second, a single norm may leave significant 'wiggle' room that leads to implicit contestation. Social norms prescribe (or proscribe) in generalities. When these generalities are translated into subjective understandings and/or actor behavior, each actor may have a slightly different understanding (as each of the agents in the Pick a Number model have a slightly different idea about the appropriate number). In a whole population, then, the intersubjective understanding will be dynamic and the intersubjective understanding will evolve. In this emergent or bottom up process, agents with subjective understandings of their intersubjective reality behave or believe in patterned, yet individually distinct ways. What we call norms are recognizable central tendencies in behaviors or expectations. In a dynamic system, however, the distribution of behaviors can change incrementally through incremental changes at the micro-level. This contestation is implicit in that it will lack a set of observable advocates arguing over the meaning of the norm, but it is still a contest in that multiple interpretations of the norm are extant in the population of agents.

Third and finally, wiggle room can lead to explicit and conscious contestation. Though a norm may be well-accepted, variants of the norm can arise in the space created by the generality of the overarching norm. What occurs then is observable contestation between variants. Contestation can reify the original, overarching norm, because that norm defines the boundaries of the debate. However, contestation can also erode the original, overarching norm as a variant may emerge that replaces the original. In addition, the policy, social, or governance outcomes of the contestation may transform the intersubjective reality, causing the original norm to 'fall.'

Each of these competitive dynamics are the target of ongoing modeling experiments. The hope is that modeling them will expand understanding of the original norm life cycle as well as advance our knowledge of norms theoretically and empirically.

5. Conclusions and Implications

This paper reports on a simple model. However, even given its simplicity, it produces complex patterns and demonstrates that norm entrepreneurs indeed can contribute to norm emergence and evolution as posited in the constructivist framework. Norm entrepreneurs alter the dynamics of a system of interacting agents, at times altering the patterns in the system toward evolving 'norms.' Further, the model provides a justification for a self-organized criticality perspective on norms. The verbal norm life cycle framework describes a driven threshold system and the results of simulating a formalized norm life cycle exhibit power law distributions as expected.

These modeling exercises have a number of significant implications. First, the model points to the conditions that govern normative dynamics, providing significant insights for empirical work on social norms. Second, the model confirms that the norm life cycle exhibits characteristics of self-organized criticality, suggesting that understanding

[6] Legro (2000) cites this as a main mechanism for norm change, identifying the need for a distinct challenger norm to arise before an extant norm will change.

norm dynamics empirically may directly entail exploring self-organized criticality. Finally, the modeling exercises provide a solid foundation for further modeling experiments, as well as empirical research into social norms.

Acknowledgements

This research was undertaken with the support of a General University Research Grant from the University of Delaware, support from National Science Foundation grant SES0083511, and a grant from the University of Delaware Research Foundation.

References

Per Bak and Kan Chen. Self-Organized Criticality. *Scientific American,* (January): 46-53, 1991.

Gregory Brunk. Why Do Societies Collapse: A Theory Based on Self-Organized Criticality. *Journal of Theoretical Politics,* 14(2), 2002.

Lars-Erik Cederman. Modeling the Size of Wars: From Billiard Balls to Sandpiles. *American Political Science Review*, 97(1): 135-150, 2003.

Lars-Erik Cederman. *Emergent Actors in World Politics*, Princeton University Press, Princeton, 1997.

Claudio Cioffi-Revilla, (ed.). *Power Laws in the Social Sciences: Discovering Complexity and Non-Equilibrium Dynamics in the Social Universe,* in preparation.

R. Conte, C. Castelfranchi. Understanding the functions of norms in social groups through simulation. *Artificial Societies: The Computer Simulation of Social Life*, N. Gilbert and R. Conte (Eds.). UCL Press, London: 252-267, 1995.

Jean Ensminger, Jack Knight. Changing Social Norms: Common Property, Bridewealth, and Clan Exogamy. *Current Anthropology,* 38 (1), 1997.

Joshua Epstein. Learning to be Thoughtless: Social Norms and Individual Computation. Santa Fe Institute Working Paper 00-03-022, 2000.

Martha Finnemore, and Kathryn Sikkink. International Norm Dynamics and Political Change. *International Organization,* 50 (4): 887-918, 1998.

Matthew Hoffmann. *Ozone Depletion and Climate Change: Constructing a Global Response,* SUNY Press, Albany, forthcoming.

Jeffrey Legro. The Transformation of Policy Ideas. *American Journal of Political Science,* 44 (3): 419-432, 2000.

James March, and Johan Olsen. *Rediscovering Institutions: the Organizational Basis of Politics*, The Free Press, New York, 1989.

Elinor Ostrom. Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives*, 14 (3): 137-158, 2000.

Nicole Saam, and Andreas Harrer. Simulating Norms, Social Inequality, and Functional Change in Artificial Societies. *Journal of Artificial Societies and Social Simulation*, 2 (1), 1999.

Mapping Deontic Operators to Abductive Expectations

Marco Alberti[⋆]
Marco Gavanelli[⋆]
Evelina Lamma[⋆]
[⋆]ENDIF - Università di Ferrara
Via Saragat, 1
44100 Ferrara, Italy
malberti@ing.unife.it
mgavanelli@ing.unife.it
elamma@ing.unife.it

Paola Mello[†]
Paolo Torroni[†]
[†]DEIS - Università di Bologna
Viale Risorgimento, 2
40136 Bologna, Italy.
pmello@deis.unibo.it
ptorroni@deis.unibo.it

Giovanni Sartor[‡]
[‡] CIRSFID - Università di Bologna
Via Galliera, 2
40100 Bologna, Italy.
sartor@cirfid.unibo.it

Abstract

A number of approaches to agent society modeling can be found in the Multi-Agent Systems literature which exploit (variants of) Deontic Logic. In this paper, after briefly mentioning related approaches, we focus on the Computational Logic (CL) approach for society modeling developed within the UE IST-2001-32530 Project (named SOCS), where obligations and prohibitions are mapped into abducible predicates (respectively, positive and negative expectations), and norms ruling the behavior of members are represented as abductive integrity constraints. We discuss how this abductive framework can deal with Deontic Logic concepts, by introducing additional integrity constraints.

1 Introduction

Several researchers have studied the concepts of norms, commitments and social relations in the context of Multi-Agent Systems (Conte et al. (1999)). Furthermore, a lot of research has been devoted in proposing architectures for developing agents with social awareness (see, for instance, Castelfranchi et al. (1999)).

Several approaches to agent society modeling have been grounded on norms and institutions (e.g., Dignum et al. (2002a,c,b); Esteva et al. (2002); Noriega and Sierra (2002)). Deontic Logic enables one to address the issue of explicitly and formally defining norms and dealing with their possible violations. It represents norms, obligations, prohibitions and permissions, and enables one to deal with predicates like "*p* ought to be done", "*p* is forbidden to be done", "*p* is permitted to be done".

In the context of the UE IST Programme, two projects (namely ALFEBIITE (ALFEBIITE) and SOCS (SOC)) have investigated the application of logic-based approaches for modeling *open*[1] societies of agents. In particular, the former focuses on the formalization of a society of agents using Deontic Logic, and the latter on a specification of an agent society which, being based on computational logic, is also executable as a verification program.

The ALFEBIITE approach (presented, for instance, by Artikis et al. (2002)) consists of a theoretical framework for providing executable specifications of particular kinds of multi-agent systems, called open computational societies, and presents a formal framework for specifying, animating and ultimately reasoning about and verifying the properties of systems where the behavior of the members and their interactions cannot be predicted in advance. Three key components of computational systems are specified, namely social constraints, social roles and social states. The specification of these concepts is based on and motivated by the formal study of legal and social systems (a goal of the ALFEBIITE

[1] For a definition of openness see Artikis et al. (2002); Hewitt (1991).

project), and therefore operators of Deontic Logic are used for expressing legal social behavior of agents (Wright (1951); van der Torre (2003)). The ALFEBIITE logical framework comprises a set of building blocks (including doxastic, deontic and praxeologic notions) as well as composite notions (including deontic right, power, trust, role and signaling acts).

The SOCS (SOC) approach to society modeling can be conceived as complementary to these efforts, since it is especially oriented toward Computational Logic aspects, and it was developed with the purpose of providing a computational framework that can be directly used for automatic verification of properties such as compliance to interaction protocols. The SOCS social model represents social norms as abductive integrity constraints, where abducibles express expectations (positive and negative) on the behavior of members of the society. The social framework is grounded on Computational Logic (CL, for short), and a declarative abductive semantics has been defined by Alberti et al. (2003a). Operationally, the application of abductive integrity constraints (named Social Integrity Constraints) by a suitable abductive proof procedure adjusts the set of social expectations as the social infrastructure acquires new knowledge from the environment in terms of happened social events. The idea of expected behavior is related, conceptually, to deontic notions such as obligation and prohibition, and it was inspired by Deontic Logic. However, in SOCS we did not exploit the full power of the standard Deontic Logic, but only abductive integrity constraints on events that are expected to happen or not to happen, and we mapped expectations into first-class abducible predicates ($\mathbf{E}$ and $\mathbf{EN}$, see the next section). Grounding the social framework on CL also smoothly provides an operational counterpart for it, in terms of an abductive proof procedure (named $\mathcal{S}$CIFF), which was obtained by extending the IFF proof procedure, proposed by Fung and Kowalski (1997).

Nonetheless, we believe that an approach grounded on CL, and abductive integrity constraints in particular, can be exploited in order to also deal with deontic concepts. This paper is meant to present a first step towards a mapping of existing formalizations of Deontic Logic onto an abductive computational framework such as SOCS'. This is achieved by means of additional (meta) integrity constraints. One of the main purposes of such mapping is to exploit the operational counterpart of the SOCS social framework (see, for instance, Alberti et al. (2004)) and the (modular) implementation of $\mathcal{S}$CIFF (suitably extended by the additional meta constraints) for the on-the-fly verification of conformance of agents to norms specified in the chosen Deontic Logic.

The paper is organized as follows. In Section 2, we briefly recall the SOCS social abductive model, and its abductive semantics. After briefly recalling Deontic Logic in Section 3, in Section 4 we show how two of its variants can be mapped into the SOCS social framework, by simply adding various (meta) integrity constraints. Section 5 briefly discusses related work. Then we conclude, and mention future work.

2 The SOCS social model

Although the SOCS project also provides a logic-based model for individual agents (see, for instance, Bracciali et al. (2004)), in this paper we abstract away from the internals of the individual agent and adopt an *external* perspective: we focus on the *observable* agent behavior, regardless of its motivation from an internal perspective. In this way, the model does not constrain the number and/or the type of agents that a society may be composed of.

The SOCS model describes knowledge about an agent society in a declarative way. Such knowledge is mainly composed of two parts: a *static* part, defining the society's organizational and "normative" elements (encoded in what we call *Social Integrity Constraints*, as we will show below), and a *dynamic* part, describing the "socially relevant" events, that have so far occurred (*happened* events). Depending on the context in which this model is instantiated, socially relevant events could indeed be physical actions or transactions, such as electronic payments. In addition to these two categories of knowledge, information about social *goals* is also maintained.

Based on the available history of events, on its specification of social integrity constraints and its goals, the society can define the social events that are expected to happen and those that are expected *not* to happen. We call these events *social expectations*; from a normative perspective, they reflect the "ideal" behavior of the agents.

2.1 Representation of the society knowledge

The knowledge in a society S is given by the following components:

- a (static) *Social Organization Knowledge Base*, denoted $SOKB$;
- a (static) set of *Social Integrity Constraints* (IC$_S$), denoted $\mathcal{IC}_S$; and

- a set of *Goals* of the society, denoted by $\mathcal{G}$.

In the following, the terms $Atom$ and $Literal$ have the usual Logic Programming meaning Lloyd (1987).

A society may evolve, as new events happen, giving rise to a sequence of society instances, each one characterized by the previous knowledge components and, in addition, a (dynamic) *Social Environment Knowledge Base*, denoted by $SEKB$.

In particular, $SEKB$ is composed of:

- *Happened events*: atoms indicated with functor $\mathbf{H}$;
- *Expectations*: events that should (but might not) happen (atoms indicated with functor $\mathbf{E}$), and events that should not (but might indeed) happen (atoms indicated with functor $\mathbf{EN}$).

In our context, "happened" events are not all the events that have actually happened, but only those observable from the outside of agents, and relevant to the society. The collection of such events is the history, $\mathbf{HAP}$, of a society instance. Events are represented as ground atoms of the form

$$\mathbf{H}(\mathit{Event}[, \mathit{Time}]).$$

For instance, in an electronic commerce context, the following atom:

$$\mathbf{H}(\mathit{tell}(a1, a2, \mathit{offer}(\mathit{scooter}, 1500), d1), 0)$$

could stand for an event about a communicative act *tell* made by agent $a1$, addressed to an agent $a2$, with subject $\mathit{offer}(\mathit{scooter}, 1500)$, at a time 0. $d1$ is, in this case, a dialogue identifier.

Expectations can be

$$\mathbf{E}(\mathit{Event}[, \mathit{Time}]) \qquad \mathbf{EN}(\mathit{Event}[, \mathit{Time}])$$

for, respectively, positive and negative expectations. $\mathbf{E}$ is a positive expectation about an event (the society expects the event to happen) and $\mathbf{EN}$ is a negative expectation, (the society expects the event not to happen[2]). Explicit negation ($\neg$) can be applied to expectations.

For instance, in an electronic commerce scenario, the following atom:

$$\mathbf{E}(\mathit{tell}(\mathit{Customer}, \mathit{Seller}, \mathit{accept}(\mathit{Item}, \mathit{Price}), \mathit{Dialogue}), T)$$

could stand for an expectation about a communicative act *tell* made by an agent ($Customer$), addressed to an agent $Seller$, with subject $accept(Item, Dialogue)$, at a time T.

The SOKB is a logic program, consisting of clauses, possibly having expectations in their body. The full syntax of SOKB is reported in Appendix.

The arguments of expectation atoms can be non-ground terms (see Alberti et al. (2003b) for a detailed discussion of variable quantification). Intuitively, variables occurring only in positive expectations are existentially quantified, whereas variables occurring only in negative expectations are universally quantified.

The following is a sample SOKB clause:

$$\mathit{on\_sale}(\mathit{Item}) \leftarrow \\ \mathbf{E}(\mathit{tell}(\mathit{Seller}, \mathit{Customer}, \mathit{offer}(\mathit{Item}, \mathit{Price}), \mathit{Dialogue}), T_0) \tag{1}$$

It says that one way to fulfill the goal: "to have a certain *item* on sale," could be to have some agent acting as a seller and offering the item at a certain price to a possible buyer.

The goal $\mathcal{G}$ of the society has the same syntax as the $Body$ of a clause in the SOKB (see Appendix), and the variables are quantified accordingly.

As an example, we can consider a society with the goal of selling items. In order to sell a scooter, the society might expect some agent to embody the role of buyer. The goal of the society could be

$$\leftarrow \mathit{on\_sale}(\mathit{scooter})$$

and the society might have, in the $SOKB$, a rule such as Eq. 1. Indeed, there could be more clauses specifying other ways of achieving the same goal.

Social Integrity Constraints are in the form of implications. The characterizing part of their syntax is reported in Appendix. For details on scope rules and quantification, see Alberti et al. (2003b). Intuitively, $\mathcal{IC}_S$ is a set of forward rules, possibly having (a conjunction of) events and expectations in their body and (a disjunction of conjunctions of) expectations in their heads. Defined predicates and Constraint Logic Programming constraints can occur in body and head, as well.

The following ic_S models one (simple) electronic vending rule, stating that each time an offer event happens, the potential buyer has to answer by accepting or refusing by a certain deadline τ.

$$\mathbf{H}(\mathit{tell}(S, B, \mathit{offer}(\mathit{Item}, \mathit{Price}), D), T0) \rightarrow \\ \mathbf{E}(\mathit{tell}(B, S, \mathit{accept}(\mathit{Item}, \mathit{Price}), D), T1), T_1 \leq T0 + \tau \vee \\ \mathbf{E}(\mathit{tell}(B, S, refuse(\mathit{Item}, \mathit{Price}), D), T1), T_1 \leq T0 + \tau$$

2.2 Abductive semantics of the Society

The SOCS social model has been interpreted in terms of Abductive Logic Programming (Kakas et al. (1998)), and an abductive semantics has been proposed for it by Alberti et al. (2003a). Abduction has

[2] $\mathbf{EN}$ is a shorthand for $\mathbf{E}$ not.

been widely recognized as a powerful mechanism for hypothetical reasoning in the presence of incomplete knowledge (Cox and Pietrzykowski (1986); Eshghi and Kowalski (1989); Kakas and Mancarella (1990); Poole (1988)).

In the SOCS social model, the idea is to exploit abduction for defining the expected behavior of the agents inhabiting the society, and an abductive proof procedure (named $\mathcal{S}$CIFF, see Alberti et al. (2003b)) to dynamically *generate* the expectations, and possibly perform the *compliance check*. By "compliance check" we mean the procedure of checking that the ic_S are not violated, together with the function of detecting fulfillment and violation of expectations.

Throughout this section, as usual when defining declarative semantics, we always consider the ground version of social knowledge base and integrity constraints, and we do not consider CLP-like constraints. Moreover, we omit the time argument in events and expectations.

First, we formalize the notions of *instance* of a society as an Abductive Logic Program (ALP, for short) Kakas et al. (1998), and *closure* of an instance. An ALP is a triple $\langle KB, \mathcal{A}, IC\rangle$ where KB is a logic program, (i.e., a set of clauses), $\mathcal{A}$ is a set of predicates that are not defined in KB and that are called *abducibles*, IC is a set of formulas called *Integrity Constraints*. An abductive explanation for a goal G is a set $\Delta \subseteq \mathcal{A}$ such that $KB \cup \Delta \models G$ and $KB \cup \Delta \models IC$, for some notion of entailment $\models$.

Definition 1 *An* instance $\mathcal{S}_{\mathbf{HAP}}$ *of a society* $\mathcal{S}$ *is represented as an ALP, i.e., a triple* $\langle P, \mathcal{E}, \mathcal{IC}_S\rangle$ *where:*

- P *is the* $SOKB$ *of* $\mathcal{S}$ *together with the history of happened events* $\mathbf{HAP}$*;*
- $\mathcal{E}$ *is the set of* abducible predicates, *namely* $\mathbf{E}$, $\mathbf{EN}$, $\neg\mathbf{E}$, $\neg\mathbf{EN}$*;*
- $\mathcal{IC}_S$ *are the social integrity constraints of* $\mathcal{S}$.

The set $\mathbf{HAP}$ characterizes the instance of a society, and represents the set of *observable* and *relevant* events for the society which have already happened. Note that we assume that such events are always ground.

A society instance is closed, when its characterizing history has been closed under the Closed World Assumption (CWA), i.e., when it is assumed that no further event will occur. In the following, we indicate a closed history by means of an overline: $\overline{\mathbf{HAP}}$.

Semantics to a society instance is given by defining those sets of expectations which, together with the society's knowledge base and the happened events, imply an instance of the goal—if any—and *satisfy* the integrity constraints.

In our definition of integrity constraint satisfaction we will rely upon a notion of entailment in a three-valued logic, it being more general and capable of dealing with both open and closed society instances. Therefore, in the following, the symbol $\models$ has to be interpreted as a notion of entailment in a three-valued setting Kunen (1987), where the history of events is open (resp. closed) for open (resp. closed) instances .

We first introduce the concept of $\mathcal{IC}_S$*-consistent set of social expectations*[3]. Intuitively, given a society instance, an $\mathcal{IC}_S$-consistent set of social expectations is a set of expectations about social events that are compatible with P (i.e., the $SOKB$ and the set $\mathbf{HAP}$), and with $\mathcal{IC}_S$.

Definition 2 ($\mathcal{IC}_S$-consistency) *Given a (closed/open) society instance* $\mathcal{S}_{\mathbf{HAP}}$*, an* $\mathcal{IC}_S$*-*consistent *set of social expectations* $\mathbf{EXP}$ *is a set of expectations such that:*

$$SOKB \cup \mathbf{HAP} \cup \mathbf{EXP} \models \mathcal{IC}_S \qquad (2)$$

(Notice that for closed instances $\mathbf{HAP}$ *has to be read* $\overline{\mathbf{HAP}}$*).*

$\mathcal{IC}_S$-consistent sets of expectations can be self-contradictory (e.g., both $\mathbf{E}(p)$ and $\neg\mathbf{E}(p)$ may belong to a $\mathcal{IC}_S$-consistent set). To avoid self-contradiction, a number of further *meta* integrity constraints have been taken into account [4]. We will show in Section 3 how these constraints, besides others, can express basic formalizations of deontic notions.

Definition 3 (E-consistency) *A set of social expectations* $\mathbf{EXP}$ *is* E-consistent *if and only if for each (ground) term* p*:*

$$\mathbf{EXP} \cup \{\mathbf{E}(p), \mathbf{EN}(p) \rightarrow false\} \not\models false \qquad (3)$$

Definition 4 ($\neg$-consistency) *A set of social expectations* $\mathbf{EXP}$ *is* $\neg$-consistent *if and only if for each (ground) term* p*:*

$$\mathbf{EXP} \cup \{\mathbf{E}(p), \neg\mathbf{E}(p) \rightarrow false\} \not\models false \qquad (4)$$

and:

$$\mathbf{EXP} \cup \{\mathbf{EN}(p), \neg\mathbf{EN}(p) \rightarrow false\} \not\models false \quad (5)$$

[3] With abuse of terminology, we call this notion $\mathcal{IC}_S$-consistency though it corresponds to the theoremhood view rather than to the consistency view defined in Fung and Kowalski (1997).

[4] In this notion, we adopt the *consistency view* defined in Fung and Kowalski (1997).

Among sets of expectations, we are interested in those satisfying Definitions 2, 3 and 4, i.e., $\mathcal{IC}_S$-, E- and ¬-consistent (we named these sets *closed*, resp. *open*, *admissible*).

Furthermore, a notion of fulfillment (similar, for positive expectations, to the notion of regimentation in Deontic Logic) was introduced in Alberti et al. (2003a), as follows.

Definition 5 (Fulfillment) *Given a (closed/open) society instance $\mathcal{S}_{\mathbf{HAP}}$, a set of social expectations* $\mathbf{EXP}$ *is* fulfilled *if and only if for all (ground) terms p:*

$$\mathbf{HAP} \cup \mathbf{EXP} \cup \{\mathbf{E}(p) \rightarrow \mathbf{H}(p)\} \cup \{\mathbf{EN}(p) \rightarrow \neg\mathbf{H}(p)\} \not\models false \tag{6}$$

Symmetrically, we define violation when the condition in Definition 5 above is not verified.

Two further notions of goal achievability and achievement were introduced in Alberti et al. (2003a) to support society goal-directed modeling. We refer to Alberti et al. (2003a) for details.

3 Deontic Notions

The birth of modern Deontic Logic can be traced back to the '50s. In the following, we only address the logical properties that are most useful in modeling legal reasoning, and norms, and refrain from addressing the logical background which provides a foundation for those properties.

Deontic Logic enables to address the issue of explicitly and formally defining norms and dealing with their possible violation. It represents norms, obligations, prohibitions and permissions, and enables one to deal with predicates like "*p* ought to be done", "*p* is forbidden to be done", "*p* is permitted to be done".

Being obligatory, being forbidden and being permitted are indeed the three fundamental *deontic statuses* of an action, upon which one can build more articulate normative conceptions. For details, refer to Sartor (2004), Chapter 15 in particular.

Obligations. To say that an action is *obligatory* is to say that the action is due, has to be held, must be performed, is mandatory or compulsory. Obligations are usually represented by formulas as:

$$\mathbf{Obl}\ A$$

where A is any (positive or negative) action description, and **Obl** is the deontic operator for obligation to be read as "it is obligatory that".

Elementary obligations can be distinguished between:

- *elementary positive obligations*, which concern positive elementary actions (e.g., "It is mandatory that John answers me");
- *elementary negative obligations*, which concern negative elementary actions (e.g., "It is mandatory that John does not smoke");

Prohibitions. The idea of obligation is paralleled with the idea of *prohibition*. Being forbidden or prohibited is the status of an action that should not be performed. In common language, and legal language as well, prohibitive propositions are expressed in various ways. For example, one may express the same idea by saying "It is forbidden that John smokes", "John must not smoke", "There is a prohibition that John smokes", and so on.

Prohibitions are usually represented by formulas as:

$$\mathbf{Forb}\ A$$

where A is any (positive or negative) action description, and **Forb** is the deontic operator for prohibition to be read as "it is forbidden that".

The notions of obligation and prohibition are logically connected, as explained in the following. Most approaches to Deontic Logic agree in assuming that, for any action A, the prohibition of A is equivalent to the obligation of omitting A:

$$\mathbf{Forb}\ A\ =\ \mathbf{Obl}\ (NON\ A) \tag{7}$$

Permissions. The third basic deontic status, besides obligations and prohibitions, is *permission*. Permissive propositions are expressed in many different ways in natural language. To express permissions in a uniform way, Deontic Logic uses the operator **Perm**. Permissions are usually represented by formulas as:

$$\mathbf{Perm}\ A$$

where A is any (positive or negative) action description, and **Perm** is the deontic operator for permission to be read as "it is permitted that".

The three basic deontic notions of obligation, prohibition and permission are logically connected. First of all, intuitively when one believes that an action is obligatory, then one can conclude that the same action is permitted.

$$\mathbf{Obl}\ A\ \ \mathbf{entails}\ \ \mathbf{Perm}\ A \tag{8}$$

Since A's obligatoriness entails A's permittedness, $\mathbf{Obl}\ A$ is incompatible with the fact that A is not permitted:

$$\mathbf{Obl}\ A\ \ \mathbf{incompatible}\ \ NON\ \mathbf{Perm}\ A \tag{9}$$

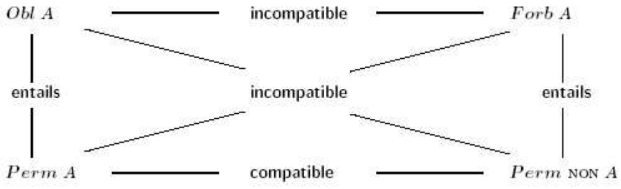


Figure 1: The first deontic square

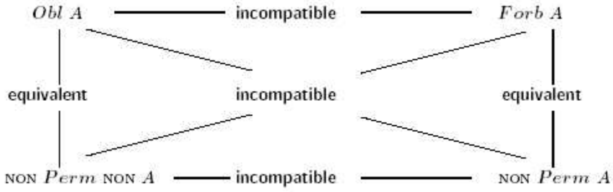


Figure 2: The second deontic square

The connection between the obligatoriness of A and the permittedness of A is replicated in the connection between the forbiddenness of A and the permittedness on A's omission: an action being forbidden entails permission to omit it, i.e.:

$$\textbf{Forb}\ A\ \ \textbf{entails}\ \ \textbf{Perm}\ NON\ A \tag{10}$$

A being forbidden entails that the omission of A is permitted. Thus, there is a contradiction between an action being forbidden and the omission of that action not being permitted.

$$\textbf{Forb}\ A\ \ \textbf{incompatible}\ \ NON\ \textbf{Perm}\ (NON\ A) \tag{11}$$

All the logical relations between deontic notions that we have just described are summarized in Figure 1. The schema shows that there is an opposition between being obliged and being prohibited: If an action A is obligatory, then its performance is permitted, which contradicts that A is forbidden.

Similarly, if an action A is forbidden, then its omission is permitted, which contradicts that A is obligatory.

It is instead compatible that both an action A is permitted and its omission $NON\ A$ also is permitted. In such a case, A would be neither obligatory nor permitted, but *facultative* (see to Sartor (2004), Chapter 15).

The deontic qualifications "obligatory" and "forbidden" are complete, in the sense that they determine the deontic status of both the action they are concerned with, and the complement of that action. In fact, on the basis of the equivalence:

$$\textbf{Obl}\ \phi = \textbf{Forb}\ NON\ \phi$$

we get the following two equivalences, the first concerning the case where ϕ is a positive action A, the second concerning the case where ϕ is the omissive action $NON\ A$ (double negations get canceled):

$$\textbf{Obl}\ A = \textbf{Forb}\ NON\ A \tag{12}$$

$$\textbf{Obl}\ NON\ A = \textbf{Forb}\ A \tag{13}$$

Of course, believing that an action is permitted amounts to believing that it is not forbidden:

$$\textbf{Perm}\ A = NON\ \textbf{Forb}\ A \tag{14}$$

This means that not being permitted amounts to being forbidden (just negate both formulas, and cancel double negations):

$$NON\ \textbf{Perm}\ A = \textbf{Forb}\ A \tag{15}$$

From this follows that an action being permitted contradicts that action being prohibited:

$$\textbf{Perm}\ A\ \ \textbf{incompatible}\ \ \textbf{Forb}\ A \tag{16}$$

Similarly, believing that an action is obligatory amounts to excluding that its omission is permitted:

$$\textbf{Obl}\ A = NON\ \textbf{Perm}\ NON\ A \tag{17}$$

Correspondingly, the obligatoriness of an action (entailing the permission to perform it) contradicts the permissiveness of its omission:

$$\textbf{Obl}\ A\ \ \textbf{incompatible}\ \ \textbf{Perm}\ NON\ A \tag{18}$$

The formulas we have just being considering are summarized in the second square of deontic notions, in Figure 2.

4 Mapping Deontic Notions onto the SOCS Social Model

This section shows how the Deontic Logic operators are mapped into SOCS social abductive model. In particular, we first show how the deontic operators can be mapped into SOCS abducible predicates standing for positive and negative expectations about social behavior (and their explicit negation). Then, we show how their logical relations can be mapped into the additional (meta) integrity constraints, considered by the (semantic and) operational machinery.

Operator	Abducibile
$\mathbf{Obl}\ A$	$\mathbf{E}(A)$
$\mathbf{Forb}\ A$	$\mathbf{EN}(A)$
$\mathbf{Perm}\ A$	$\neg\mathbf{EN}(A)$
$\mathbf{Perm}\ NONA$	$\neg\mathbf{E}(A)$

Table 1: Deontic notions as expectations

4.1 Mapping deontic operators onto expectations

Conceptually, a natural correspondence appears between the notion of obligation (which requires an action to be performed) and ours of positive expectation (which requires an event to belong to the history in order to achieve fulfillment, as of Def. 5). In the same way, a negative expectation corresponds to a prohibition. Moreover, since a negative expectation $\mathbf{EN}(A)$ has to be read as *it is expected not A* (i.e., it is a shorthand for $\mathbf{E}(not\ A)$), its (explicit) negation, $\neg\mathbf{EN}(A)$, corresponds to permission of A.

Therefore, the three deontic notions can be mapped into expectations as summarized by the first three lines in Table 1.

Furthermore, due to the logical relations among obligation, prohibition and permission discussed in Section 3, the fourth line of Table 1 shows how to map permission of a negative action. Notice that, while both NON and $\neg$ represent the explicit negation of their argument, we keep the different symbols for uniformity with the original contexts.

It is worth noticing, however, that despite this natural mapping the deontic notions and SOCS social expectations are grounded on different semantic approaches, inherited from modal logic the former, and based on abduction the latter.

4.2 Logical relations among deontic operators as abductive integrity constraints

Let us first consider the relations summarized in the second square of deontic notions, in Figure 2. By adopting the mapping summarized in Table 1, the equivalence relations straightforwardly arise from the uniform treatment of symbols NON, $\neg$ and not, and from their idempotency.

Incompatibility relations summarized in Figure 2 emerge between the notion of obligation and prohibition (horizontal arc), and, respectively, between obligation and permission of opposite, and prohibition and non permission of opposite (diagonal arcs). By adopting the mapping summarized in Table 1, the first incompatibility is captured by SOCS social abductive semantics into the notion of E-consistency (Definition 3), i.e., by requiring that, for each A, the addition to the expectation set of the integrity constraint:

$$\mathbf{E}(A), \mathbf{EN}(A) \rightarrow false$$

does not lead to inconsistency.

The latter two incompatibilities (corresponding to diagonal arcs in Table 1) are captured, instead, by the notion of $\neg$-consistency (Definition 4), i.e., by requiring that, for each A, the addition to the expectation set of the integrity constraints:

$$\mathbf{E}(A), \neg\mathbf{E}(A) \rightarrow false$$

and

$$\mathbf{EN}(A), \neg\mathbf{EN}(A) \rightarrow false$$

does not lead to inconsistency.

The notions of *E*-consistency and $\neg$-consistency (and associated integrity constraints) also correspond to incompatibility relations in the first square of deontic notions, in Figure 1.

Furthermore, the two entailment relations occurring in the first square can be captured by considering additional integrity constraints (possibly added to the set $\mathcal{IC}_S$), relating positive and negative expectations as follows:

$$\mathbf{E}(A) \rightarrow \neg\mathbf{EN}(A)$$

and

$$\mathbf{EN}(A) \rightarrow \neg\mathbf{E}(A)$$

In practice, these two constraints, when added to $\mathcal{IC}_S$ and therefore considered in $\mathcal{IC}_S$-consistency, enforce the set of expectations to be “completed”, i.e., for each positive expectation $\mathbf{E}(A)$ the explicit negation of its negative counterpart, $\neg\mathbf{EN}(A)$ had to be included in the expectation set (in order to get its admissibility), and for each negative expectation $\mathbf{EN}(A)$ the explicit negation of its positive counterpart, $\neg\mathbf{E}(A)$ had to be included as well.

Finally, a notion of *regimentation* can be considered too, by enforcing obligatory actions to happen and prohibited actions not to happen. This can be easily obtained by adding to the $\mathcal{IC}_S$ the following two integrity constraints, mapping positive/negative expectations into positive/negative events:

$$\mathbf{E}(A) \rightarrow \mathbf{H}(A)$$

and

$$\mathbf{EN}(A) \rightarrow \neg\mathbf{H}(A)$$

Notice that these two conditions correspond to the (meta) integrity constraints required for fulfillment of expectation sets (see Definition 5). The adopted notion of fulfillment in the declarative semantics, however, just test that these two constraints are not violated (by adopting the consistency view discussed by Fung and Kowalski (1997)), whereas if we add them to the set $\mathcal{IC}_S$ the $\mathcal{IC}_S$-consistency test (by adopting the theoremhood view, also discussed by Fung and Kowalski (1997)) would exploit them to also make events happening or not in the social environment.

A notable difference, from the representation point of view, is that in SOCS social integrity constraints can only express disjunctions of expectations, such that $\mathbf{E}(A) \vee \mathbf{E}(B)$ (which expresses that at least one of the two between A and B events is expected). In Deontic Logic, instead, one usually expresses the obligatoriness of disjunctions, i.e., $\mathbf{Obl}(A \vee B)$. In Kripke-like semantics (adopted for Deontic Logic), however, this is not equivalent to state $\mathbf{Obl}(A) \vee \mathbf{Obl}(B)$ [5].

The SOCS formalism based on $\mathcal{IC}_S$ constraints can capture, instead, in a computational setting, the concept of (conditional) obligation with deadline presented by Dignum et al. (2002a), with an explicit mapping of time. Dignum *et al.* write: `Oa(r<d|p)` to state that if the precondition `p` becomes valid, the obligation becomes active. The obligation expresses the fact that `a` is expected to bring about the truth of `r` before a certain condition `d` holds.

For instance, if we have:

$$\begin{aligned} &p = \mathbf{H}(tell(S, a, request(G), D, T)) \\ &r = \mathbf{H}(tell(a, S, answer(G), D, T')), T' > T \\ &d = T' > T + 2 \end{aligned}$$

we can map `Oa(r<d|p)` into a ic_S:

$$\mathbf{H}(tell(S, a, request(G), D), T) \rightarrow \\ \quad \mathbf{E}(tell(a, S, answer(G), D), T'), T' > T, T' \leq T + 2.$$

5 Related Work

There exist a number of approaches based on Deontic Logic to formally defining norms and dealing with their possible violations.

Among the organizational models, Dignum et al. (2002a,c,b) exploit Deontic Logic to specify the society norms and rules. Their model is based on a framework which consists of three interrelated models: organizational, social and interaction. The *organizational model* defines the coordination and normative elements and describes the expected behavior of the society. Its components are roles, constraints, interaction rules, and communicative and ontology framework. The *social model* specifies the contracts that make explicit the commitments regulating the enactment of roles by individual agents. Finally, the *interaction model* describes the possible interactions between agents by specifying contracts in terms of description of agreements, rules, conditions and sanctions.

The reduction of deontic concepts such as obligations and prohibitions has been the subject of several past works: notably, by Anderson (1958) (according to which, informally, A is obligatory iff its absence produces a state of violation) and by Meyer (1988) (where, informally, an action A is prohibited iff its being performed produces a state of violation). These two reductions strongly resemble our definition of fulfillment (Def. 5), which requires positive (resp. negative) expectations to have (resp. not to have) a corresponding event.

van der Torre and Tan (1999) show the relation between diagnostic reasoning and deontic logic, importing the *principle of parsimony* from diagnostic reasoning into their deontic system, in the form of a requirement to minimize the number of violations. The management of violations (minimizing their number and possibly recovering from them) is currently not addressed by the SOCS framework and is subject of future work.

Boella and van der Torre (2003) discuss how a normative system can be seen as a normative agent, equipped with mental attitudes, about which other agents can reason. The social infrastructure in the SOCS model could be viewed as an agent whose knowledge base is the society specification, and whose reasoning process is the $\mathcal{S}$CIFF proof procedure.

Deontic operators have been used not only at the social level, but also at the agent level. Notably, in IMPACT (Arisha et al. (1999); Eiter et al. (1999)), agent programs may be used to specify what an agent is obliged to do, what an agent may do, and what an agent cannot do on the basis of deontic operators of Permission, Obligation and Prohibition (whose semantics does not rely on a Deontic Logic semantics). In this respect, the IMPACT and SOCS social models have similarities even if their purpose and expressivity are different. The main difference is that the goal of agent programs in IMPACT is to express and determine by its application the behavior of a single

[5] The two possible worlds $(A \wedge NONB)$ and $(NONA \wedge B)$ satisfy $\mathbf{Obl}(A \vee B)$, but not $\mathbf{Obl}(A) \vee \mathbf{Obl}(B)$.

agent, whereas the SOCS social model goal is to express rules of interaction and norms, that instead cannot really determine and constrain the behavior of the single agents participating to a society, since agents are autonomous.

6 Conclusion and Future Work

In this work, we have discussed how the Computational Logic-based framework for modeling societies of agents developed within the UE IST-2001-32530 project (named SOCS) can be exploited to express different variants of Deontic Logic. SOCS approach for modeling open societies is based on an abductive framework, where obligations and prohibitions are mapped into abducible predicates (respectively, positive and negative expectations), and norms ruling the behavior of members are represented as abductive integrity constraints. The SOCS social abductive framework can easily express different Deontic Logics, by means of additional (meta) integrity constraints.

This mapping is relevant from the representation point of view, but this is even more interesting from the computational viewpoint. In fact, since SOCS abductive social model is grounded on Computational Logic, it also offers an operational counterpart as an abductive proof procedure named $\mathcal{S}$CIFF which extends the IFF proof procedure by Fung and Kowalski (1997). $\mathcal{S}$CIFF is based on transitions able to deal with dynamic events, propagate social integrity constraints, etc., and it was proved sound with respect to the defined abductive declarative semantics. In particular, $\mathcal{S}$CIFF is able to verify the conformance of agent interactions with respect to the specified norms as $\mathcal{IC}_S$. Its implementation (see Alberti et al. (2004)) has been obtained in SICStus Prolog (SICStus), by exploiting the *Constraint Handling Rules* (CHR) library (Frühwirth (1998)). Both $\mathcal{S}$CIFF transitions and the meta integrity constraints (for E- and $\neg$-consistency) have been mapped into CHR rewriting rules. This modular implementation can be easily extended by considering the additional integrity constraints defined in this paper, in order to deal with the different variants of Deontic Logic discussed. This is subject for future work.

Acknowledgments

This work has been supported by the European Commission within the SOCS project (IST-2001-32530), funded within the Global Computing Programme and by the MIUR COFIN 2003 projects *La Gestione e la negoziazione automatica dei diritti sulle opere dell'ingegno digitali: aspetti giuridici e informatici* and *Sviluppo e verifica di sistemi multiagente basati sulla logica*.

References

Societies Of ComputeeS (SOCS): a computational logic model for the description, analysis and verification of global and open societies of heterogeneous computees. `http://lia.deis.unibo.it/Research/SOCS/`.

M. Alberti, M. Gavanelli, E. Lamma, P. Mello, and P. Torroni. An Abductive Interpretation for Open Societies. In A. Cappelli and F. Turini, editors, *AI*IA 2003: Advances in Artificial Intelligence, Proceedings of the 8th Congress of the Italian Association for Artificial Intelligence, Pisa*, volume 2829 of *Lecture Notes in Artificial Intelligence*, pages 287–299. Springer-Verlag, September 23–26 2003a. `http://www-aiia2003.di.unipi.it`.

Marco Alberti, Federico Chesani, Marco Gavanelli, Evelina Lamma, Paola Mello, and Paolo Torroni. Compliance verification of agent interaction: a logic-based tool. pages 570–575, Vienna, Austria, April 13-16 2004. Austrian Society for Cybernetic Studies.

Marco Alberti, Marco Gavanelli, Evelina Lamma, Paola Mello, and Paolo Torroni. Specification and verification of interaction protocols: a computational logic approach based on abduction. Technical Report CS-2003-03, Dipartimento di Ingegneria di Ferrara, Ferrara, Italy, 2003b. Available at `http://www.ing.unife.it/aree_ricerca/informazione/cs/technical_reports`.

ALFEBIITE. ALFEBIITE: A Logical Framework for Ethical Behaviour between Infohabitants in the Information Trading Economy of the universal information ecosystem. IST-1999-10298, 1999. Home Page: `http://www.iis.ee.ic.ac.uk/~alfebiite/ab-home.htm`.

A. Anderson. A reduction of deontic logic to alethic modal logic. *Mind*, 67:100–103, 1958.

K. A. Arisha, F. Ozcan, R. Ross, V. S. Subrahmanian, T. Eiter, and S. Kraus. IMPACT: a Platform for Collaborating Agents. *IEEE Intelligent Systems*, 14(2):64–72, March/April 1999.

A. Artikis, J. Pitt, and M. Sergot. Animated specifications of computational societies. In C. Castelfranchi and W. Lewis Johnson, editors, *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2002), Part III*, pages 1053–1061, Bologna, Italy, July 15–19 2002. ACM Press. ISBN 1-58113-480-0. http://portal.acm.org/ft_gateway.cfm?id=545070&type=pdf&dl=GUIDE&dl=ACM%&CFID=4415868&CFTOKEN=57395936.

Guido Boella and Leendert W. N. van der Torre. Attributing mental attitudes to normative systems. In *AAMAS*, pages 942–943, 2003.

Andrea Bracciali, Neophytos Demetriou, Ulle Endriss, Antonis Kakas, Wenjin Lu, Paolo Mancarella, Fariba Sadri, Kostas Stathis, Francesca Toni, and Giacomo Terreni. The KGP model of agency: Computational model and prototype implementation. In *Global Computing Workshop, Rovereto, Italy, March 2004*, Lecture Notes in Artificial Intelligence. Springer-Verlag, 2004. to appear.

C. Castelfranchi, F. Dignum, C.M. Jonker, and J. Treur. Deliberative normative agents: Principles and architecture. In Nicholas R. Jennings and Yves Lespérance, editors, *Intelligent Agents VI, Agent Theories, Architectures, and Languages, 6th International Workshop, ATAL '99, Orlando, Florida, USA, Proceedings*, number 1757 in Lecture Notes in Computer Science, pages 364–378. Springer-Verlag, 1999.

R. Conte, R. Falcone, and G. Sartor. Special issue on agents and norms. *Artificial Intelligence and Law*, 1(7), March 1999.

P. T. Cox and T. Pietrzykowski. Causes for events: Their computation and applications. In *Proceedings CADE-86*, pages 608–621, 1986.

V. Dignum, J. J. Meyer, F. Dignum, and H. Weigand. Formal specification of interaction in agent societies. In *Proceedings of the Second Goddard Workshop on Formal Approaches to Agent-Based Systems (FAABS), Maryland*, October 2002a.

V. Dignum, J. J. Meyer, and H. Weigand. Towards an organizational model for agent societies using contracts. In C. Castelfranchi and W. Lewis Johnson, editors, *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2002), Part II*, pages 694–695, Bologna, Italy, July 15–19 2002b. ACM Press. ISBN 1-58113-480-0. http://portal.acm.org/ft_gateway.cfm?id=544909&type=pdf&dl=GUIDE&dl=ACM%&CFID=4415868&CFTOKEN=57395936.

V. Dignum, J. J. Meyer, H. Weigand, and F. Dignum. An organizational-oriented model for agent societies. In *Proceedings of International Workshop on Regulated Agent-Based Social Systems: Theories and Applications. AAMAS'02, Bologna*, 2002c.

T. Eiter, V.S. Subrahmanian, and G. Pick. Heterogeneous active agents, I: Semantics. *Artificial Intelligence*, 108 (1-2):179–255, March 1999.

K. Eshghi and R. A. Kowalski. Abduction compared with negation by failure. In G. Levi and M. Martelli, editors, *Proceedings of the 6th International Conference on Logic Programming*, pages 234–255. MIT Press, 1989.

M. Esteva, D. de la Cruz, and C. Sierra. ISLANDER: an electronic institutions editor. In C. Castelfranchi and W. Lewis Johnson, editors, *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2002), Part III*, pages 1045–1052, Bologna, Italy, July 15–19 2002. ACM Press. ISBN 1-58113-480-0. http://portal.acm.org/ft_gateway.cfm?id=545069&type=pdf&dl=GUIDE&dl=ACM%&CFID=4415868&CFTOKEN=57395936.

T. Frühwirth. Theory and practice of constraint handling rules. *Journal of Logic Programming*, 37(1-3):95–138, October 1998.

T. H. Fung and R. A. Kowalski. The IFF proof procedure for abductive logic programming. *Journal of Logic Programming*, 33(2):151–165, November 1997. ISSN 0743-1066.

C. Hewitt. Open information systems semantics for distributed artificial intelligence. *Artificial Intelligence*, 47 (1-3):79–106, 1991.

A. C. Kakas, R. A. Kowalski, and F. Toni. The role of abduction in logic programming. In D. M. Gabbay, C. J. Hogger, and J. A. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 5, pages 235–324. Oxford University Press, 1998.

A. C. Kakas and P. Mancarella. On the relation between Truth Maintenance and Abduction. In T. Fukumura, editor, *Proceedings of the 1st Pacific Rim International Conference on Artificial Intelligence, PRICAI-90, Nagoya, Japan*, pages 438–443. Ohmsha Ltd., 1990.

K. Kunen. Negation in logic programming. In *Journal of Logic Programming*, volume 4, pages 289–308, 1987.

J. W. Lloyd. *Foundations of Logic Programming*. Springer-Verlag, 2nd extended edition, 1987. ISBN 3-540-18199-7.

J. J. Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame J. of Formal Logic*, 29(1):109–136, 1988.

P. Noriega and C. Sierra. Institutions in perspective: An extended abstract. In *Sixth International Workshop CIA-2002 on Cooperative Information Agents*, volume 2446 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2002.

D. L. Poole. A logical framework for default reasoning. *Artificial Intelligence*, 36(1):27–47, 1988.

Giovanni Sartor. *Legal Reasoning*, volume 5 of *Treatise*. Kluwer, Dordrecht, 2004.

SICStus. SICStus prolog user manual, release 3.11.0, October 2003. `http://www.sics.se/isl/sicstus/`.

L. van der Torre. Contextual deontic logic: Normative agents, violations and independence. *Annals of Mathematics and Artificial Intelligence*, 37(1):33–63, 2003.

Leendert W. N. van der Torre and Yao-Hua Tan. Diagnosis and decision making in normative reasoning. *Artif. Intell. Law*, 7(1):51–67, 1999.

G.H. Wright. Deontic logic. *Mind*, 60:1–15, 1951.

Appendix

The SOKB is a logic program, consisting of clauses, possibly having expectations in their body. The full syntax of SOKB is the following:

$$\begin{array}{rcl} Clause & ::= & Atom \leftarrow Body \\ Body & ::= & ExtLiteral\ [\ \wedge ExtLiteral\ ]^{\star} \\ ExtLiteral & ::= & Literal \mid Expectation \mid Constraint \\ Expectation & ::= & [\neg]\mathbf{E}(Event\ [,T]) \mid [\neg]\mathbf{EN}(Event\ [,T]) \end{array} \tag{19}$$

Social Integrity Constraints are in the form of implications. The characterizing part of their syntax is the following:

$$\begin{array}{rcl} ic_S & ::= & \chi \rightarrow \phi \\ \chi & ::= & (HEvent|Expectation)\ [\wedge BodyLiteral]^{\star} \\ BodyLiteral & ::= & HEvent|Expectation|Literal|Constraint \\ \phi & ::= & HeadDisjunct\ [\ \vee HeadDisjunct\ ]^{\star}|\bot \\ HeadDisjunct & ::= & Expectation\ [\ \wedge (Expectation|Constraint)]^{\star} \\ Expectation & ::= & [\neg]\mathbf{E}(Event\ [,T]) \mid [\neg]\mathbf{EN}(Event\ [,T]) \\ HEvent & ::= & [\neg]\mathbf{H}(Event\ [,T]) \end{array} \tag{20}$$

Given an ic_S $\chi \rightarrow \phi$, χ is called the *body* (or the *condition*) and ϕ is called the *head* (or the *conclusion*).

For details on scope rules and quantification, please refer to Alberti et al. (2003b).

Towards Norm-Governed Self-Organising Networks

Lloyd Kamara⋆ Jeremy Pitt⋆ Marek Sergot†
⋆Intelligent Systems and Networks Group, EEE Dept., Imperial College London
{l.kamara,j.pitt}@imperial.ac.uk
Tel.: +44 (0)20 7594 6187 Fax: +44 (0)20 7594 6274
†Department of Computing, Imperial College London
m.sergot@imperial.ac.uk
Tel.: +44 (0)20 7594 8218 Fax: +44 (0)20 7581 8024

Abstract

Ad hoc networks may be viewed as computational systems whose members may fail to, or choose not to, comply with the rules governing participation. From this perspective, there is a need for mechanisms to model, monitor and manage interactions in these networks in order to promote their smooth running and correct operation. We propose a norm-governed approach to satisfy these requirements, comprising an agent architecture with an objective reasoning capacity (allowing agents to reason about normative positions) and a suite of protocols for network management. In this paper, we describe a corresponding system architecture that, if successful, will lead to ad hoc networks capable of demonstrating self-organising behaviour in accordance with external system and normative specifications.

1 Introduction

Ad hoc network(s) (AHN) are chiefly characterised by a dynamic, resilient topology and a self-organising capacity (Perkins, 2001). An AHN is typically based on wireless technology and may be short-lived, oriented toward spontaneous rather than long-term interoperation. Such a network may be formed, for example, by devices owned by participants in a workshop or project meeting (for sharing and co-authoring documents); by consumers entering and leaving an 802.11 wireless hot spot covering a shopping mall (for trading goods in a C2C-style framework wherein potential buyers are matched with sellers); or by emergency or disaster relief workers (in the absence of the usual static support infrastructure) (Wu and Stojmenovic, 2004).

AHN have significant technological and management requirements due to the potential number and variability of interactions that take place within them. A key management issue is that of resource sharing: how to effectively conserve or employ collective resources (such as bandwidth, power, processor cycles and file storage) *logically* given their discontinuous distribution across the network *physically* and their respective ownership by individual participants. The aim of our present research is to investigate to what extent such issues can be addressed by viewing AHN as instances of *normative systems* (Jones and Sergot, 1993). A normative system can be considered a collection of recognised (authoritative) rights or standards that can be used to regulate the behaviour of those to whom the system applies, either through subscription (voluntary) or conscription (obligatory). From this, we derive the concept of a *norm-governed (agent) system* — a multi-agent system or society in which the members' behaviour and interactions are defined (or characterised) by a normative system. The term *norm* in this context covers permissions, obligations and other more complex relations that may exist between agents and their societies. It is meaningful to identify such concepts in agent societies where the behaviour of members can deviate from the optimal.

We have identified two levels of abstraction for examining AHN from a norm-governed perspective: a *physical level* and an *application level*. The physical level concerns fundamental network services — such as registration and routing — that manage the dynamic collection of network *nodes* constituting an AHN. The application level uses the logical, operational platform provided by the physical level to conduct higher-level tasks in a manner largely independent of the network's physical topology. We believe that at this level, an AHN can be viewed as an *open agent society* (Artikis, 2003; Artikis et al., 2003) (OAS) — that is, a computational (agent) community exhibiting the following characteristics:

- *agents are owned and operated by different par-*

ties and therefore have varied characteristics (architectures) and internal behaviours (such as motivations).

- *the internal state of agents can only be inferred* as each agent architecture is of arbitrary design and not externally accessible.

- *the concept of global (communal) utility is optional* thus agents may be individually motivated to pursue goals at odds with altruistic (community) endeavour.

At the physical level of an AHN, it is probable that system components will fail to behave as they ought to — not from willfulness or to seek advantage over others but simply because of the inherently transient nature of the AHN. It is therefore meaningful to speak of system components failing to comply with their obligations, of permitted/forbidden actions, and even of sanctions (though clearly not of punishments).

In the rest of this paper, we further illustrate the motivation for a norm-governed, open-agent society approach to AHN self-organisation and management. We first present an AHN model and compare some of its key characteristics to those of an open agent society. We then demonstrate the principles behind our previous work on specifying and modelling open agent societies. We subsequently outline how those principles may be effectively applied in the AHN context. We also outline a norm-governed agent architecture proposal based on a combination of the concepts in the preceding sections and discuss the protocols such agents would use in a norm-governed AHN. We compare our proposal to existing, related work and finally conclude with a summary of the presented material and future work prospects.

2 Modelling ad hoc networks

An AHN comprises a number of interacting nodes whose individual *communications properties*, *resources* and *location* are combined to define behaviour at the physical level. A node's communications might include, for example, transmission strength and reception sensitivity. The basic node structure is depicted in Figure 1 (left-hand side), where the circle surrounding a node denotes its communications range. The communications ranges of nodes within a typical AHN do not wholly coincide: there may be nodes in the AHN that cannot directly interact. For this reason, AHN employ a decentralised communications model in which traffic can be routed by intermediary nodes if a direct route is unavailable or undesirable. In the example AHN of Figure 1 (right-hand side), direct communication is possible only where circles — and hence communications ranges — overlap. The circles are also shaded differently to denote distinct bandwidth characteristics: a darker shading denotes greater bandwidth. This emphasises the potentially heterogeneous nature of the devices (nodes) comprising the network. Note that for the sake of simplicity we do not distinguish between node reception and transmission ranges but treat the two as the same. In reality, a node's reception range will typically exceed its transmission range.

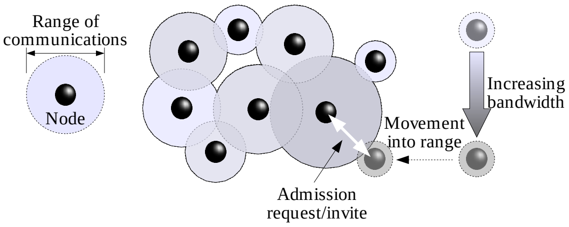


Figure 1: AHN node representation (left) and a corresponding AHN

While joining (or forming) an AHN always takes place in a predictable manner, the same does not apply to 'leaving'. The nodes of a typical AHN are mobile and may therefore go out of range, temporarily or otherwise. Scope exists at both physical and application levels to make arrangements for these communications interruptions. Example mechanisms include place-holding, message forwarding, automated workload redistribution and redundancy strategies.

The similarity between AHN and open agent societies can be observed in relation to the above model. Both are dynamic constructs, exhibiting *decentralised command and control*, with nodes/agents performing *local decision-making*. Both offer *open access*; subject to following the rules/protocols governing basic participation, others can attempt to join an existing AHN/society at will. Both involve *delegated responsibility*, since agents and nodes operate on behalf of their owners, implying the concept of *accountability*, wherein the actions of a node or agent may have legal consequences for its operator or owner. This is particularly the case in *institutionalised interactions*, such as commercial transactions. Both AHN and open agent societies allow the possibility of processes deviating from optimal behaviour — such as failing to conform to a specification or making wrong decisions. The nodes of an AHN can therefore be as heterogeneous, opaque and non-compliant as their counterparts within an open agent society are allowed to be. In the following section, we demonstrate how agent behaviour in such societies can be specified, modelled and ultimately constrained within a norm-governed framework.

3 Specifying and modelling open agent societies

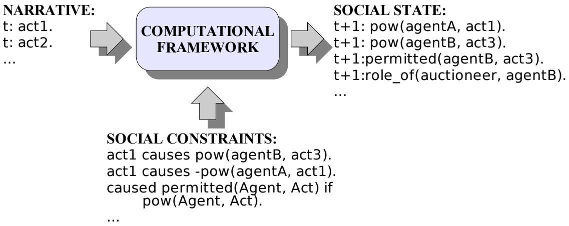
In previous work (Artikis, 2003; Artikis et al., 2003), we have used the action language C^+ (Giunchiglia et al., 2003) and the Event Calculus (Kowalski and Sergot, 1986) to represent the normative positions that arise when an open agent society is viewed as a norm-governed system. These languages provide a means to express the *social constraints*, or laws, governing an agent society — for example, what kind of actions 'count as' valid (alternatively 'effective' or 'well-formed') actions. An action counts as a valid one if the agent that performs it has the *institutionalised power* (Jones and Sergot, 1996) to do so.

Institutional (or institutionalised) power refers to a standard feature of all norm-governed organisations whereby designated agents are empowered, by the institution, to create facts that have a conventional significance within that institution. The concept is seen in (Searle, 1971), which identifies *brute facts* (observable facts associated with the physical aspect of a system) and *institutional facts* (a type of social fact defined within an institution). If node A is within physical communications range of node B, for instance, this would be said to be a brute fact. An example of an institutional fact would be if node B were said to occupy the role of communications proxy between node A and node C. Jones and Sergot (1996) present a formalisation of institutionalised power in terms of an even more fundamental notion; namely that within a given institution, certain kinds of acts or states of affairs have conventional significance, in that they *count as* other kinds of acts or states of affairs.

Consider an AHN formed for the purpose of task allocation and performance via the *contract net protocol* (Smith and Davis, 1978). In this context, one node occupies the role of *manager*, while other nodes are potential *contractors*. If the manager performs the speech act 'contract *cid* has been awarded to contractor C', this *counts as*, in the AHN institution, a means of establishing that contractor C has indeed been awarded contract *cid*. The manager is said to have had the power (or been empowered) within the AHN institution, to establish that the contract is awarded. Were another node to attempt the same action, there would be no effect. We use the term *valid* to denote institutionally empowered action performance. It should also be noted that the act of awarding the contract results in further changes to the powers of both the manager and the relevant contractor.

Distinguishing between valid and invalid actions enables the separation of 'meaningful' from 'meaningless' activities. We are similarly able to specify what kind of actions are permitted. Determining the permitted, prohibited and obligatory actions enables the classification of agent behaviour as legal or illegal, ethical or unethical, social or anti-social, and so on. This classification can form the basis of sanctions and enforcement policies to handle non-compliant behaviour.

In addition to enabling us to represent the institutionalised powers of open agent society members, our computational framework allows us to formally represent and reason about the actions of those members. Figure 2 illustrates the framework being used in a *predictive* capacity. Given a delimited history of actions and events (a *narrative*), we can use the computational framework to interpret the social constraints and determine possible future states of the society. In Figure 2, a *Prolog*-like syntax is used to express the actions and events within the narrative and are of the form t: act1 to denote action act1 having taken place at time t. The social constraints contain causation rules: for example, act1 causes pow(agentB, act3) signifies that when act1 is performed, agent agentB subsequently has the institutional power to perform act3. Powers can also be revoked by actions and events, hence the expression act1 causes $-$pow(agentA, act1). The relationship between power and permission may be application-dependent, as well as being institution-specific: here, the final social constraint establishes that any institutionally empowered action is permitted (note that capitalised terms are variables ranging over agents and actions accordingly). The social state that results from applying the computational framework to the narrative and social constraints shows what will be true of future time-points (in this example, t+1).



```
NARRATIVE:
t: act1.
t: act2.
...
```

```
SOCIAL STATE:
t+1: pow(agentA, act1).
t+1: pow(agentB, act3).
t+1:permitted(agentB, act3).
t+1:role_of(auctioneer, agentB).
...
```

```
SOCIAL CONSTRAINTS:
act1 causes pow(agentB, act3).
act1 causes -pow(agentA, act1).
caused permitted(Agent, Act) if
        pow(Agent, Act).
...
```

Figure 2: A framework for predicting social state in open agent societies

The computational framework can also be employed in a *planning* capacity: given an initial state, desired final state and set of social constraints, it can infer the required actions and events to get from desired to initial state (Figure 3). In both prediction and planning usage, the social constraints serve as a form of *executable specification* (Artikis, 2003). The specification can be used at the macro-level — for example, by a society monitor, which checks, from the externally observable events, whether agents comply with regulations. The specification may also be used at the micro-level, if some or all of the participating agents have an *objective reasoning* capability — that is, they can compute prediction and planning queries themselves.

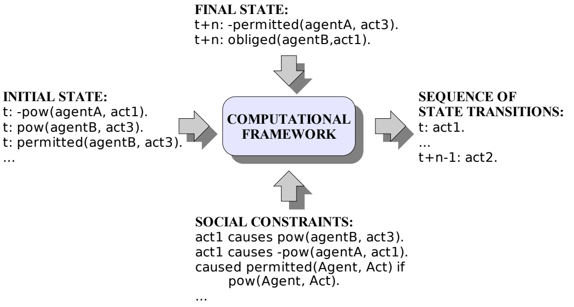

FINAL STATE:
t+n: -permitted(agentA, act3).
t+n: obliged(agentB,act1).
INITIAL STATE:
t: -pow(agentA, act1).
t: pow(agentB, act3).
t: permitted(agentB, act3).
...
COMPUTATIONAL FRAMEWORK
SEQUENCE OF STATE TRANSITIONS:
t: act1.
...
t+n-1: act2.
SOCIAL CONSTRAINTS:
act1 causes pow(agentB, act3).
act1 causes -pow(agentA, act1).
caused permitted(Agent, Act) if
pow(Agent, Act).
...

Figure 3: A planning application of the computational framework

We maintain that any networked operation having a dynamic topology and featuring distributed functionality, decentralised command and control, local decision-making and semi-autonomous action can be regarded as an instance of an open agent society. The basic AHN model of Section 2 falls into this category. We therefore believe that the same norm-governed approach to the specification and management of open agent societies can be applied in the AHN context. This position motivates the use of objective reasoning in an AHN at institutional and individual node levels. We thus introduce two new elements to the management and organisation of AHN: the use of normative concepts at an institutional level and the potential influence arising from awareness of normative positions at the node level. In the following sections, we give further details of this strategy.

4 Norm-governed ad hoc network management

We seek to address AHN management issues such as scale, complexity and resource-sharing by exploiting the analogy between AHN and open agent societies. To do so, we first equate a node within an AHN to an agent within a society. We then envisage the agents being able to represent normative concepts, like permission, obligation and institutional power. These representations alone, however, are insufficient for our purposes: the agents must be able to reason about them as well. To this end, we propose an *agent architecture* that incorporates the representations into its decision-making process, through the objective reasoning capability described in Section 3. While this suggestion may at first appear to run counter to the principles of open agent societies (specifically the inaccessibility of an agent's internal state), we argue that it is acceptable within a controlled simulation environment. That is, as designers and experimenters, we may choose to model an AHN wherein the participating nodes are based upon a single agent architecture, but exhibit different characteristics through instantiation and parameterisation. Additionally, the agents would not necessarily be 'aware' of their architectural commonality: heterogeneity is therefore maintained from an internal perspective Kamara et al. (2003).

The social constraints mentioned previously serve here to define an AHN's *mores*, covering procedures like network formation, participation, ejection and disbanding. They may also be thought of as establishing a contract between participants and the society, one whose terms and conditions define the basic nature of interactions.

AHN — in particular, those based on wireless technologies — are susceptible to disruption from participating nodes moving out of communications range. Such departures may be unintentional but within a norm-governed context, they can take on additional significance. For example, if a node is perceived to have stopped responding to access requests for a shared resource under its control, this may be interpreted as a violation of a norm. The underlying problem, is one that can arise in most distributed computing environments. Time-out mechanisms can be introduced to counter this, but the potential remains for change in the environment to be misclassified as agent (in)action. The problem lies in the underlying physical configuration of the AHN. As such, we acknowledge its existence but do not overly focus upon it here.

We have described previously how objective reasoning could allow an agent to determine significant normative positions that arise during its transactions within an agent society. To use this feature effectively, an agent must possess the cognitive capability to factor norms into its broader decision-making processes. There is an additional need for appropriate coordination mechanisms that enable agents to interact in such a way that allows norms to arise and be recognised within an AHN. In the following sections, we outline an approach to incorporating the concepts described above within an agent architecture and discuss the protocols that guide the interactions of peers.

5 A norm-aware agent architecture for AHN

Figure 4 provides an overview of an architecture allowing agents to be aware of the social constraints governing their behaviour, and to consider such constraints in their decision-making processes. The architecture augments the traditional logical groupings of control and state by a *social state* component and the inclusion of (or link to) an *objective reasoning module* (ORM). By *inclusion* we mean the case in which the agent has an inherent capacity to reason about social constraints. The alternative, as mentioned previously, is for the agent to *link* (via an API), to an external module or process which is able to reason about social constraints on the invoking agent's behalf.

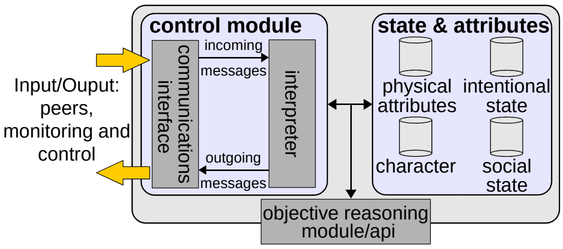


Figure 4: Inside a norm-governed agent

Given a *narrative* (the events witnessed by a participant P) the ORM of P computes the instantaneous social state (the institutional powers, permissions, obligations and sanctions associated with P) and, possibly, some of its peers, for any given point in time. The computation of these normative relations is based on the social constraints of the norm-governed AHN (ngAHN). That is, an agent carries out actions and witnesses events, the occurrence of which, when combined with the social constraints and interpreted by the ORM, dictate what changes (if any) occur in that agent's powers, permissions and obligations. By invoking or consulting the ORM, the agent interpreter can determine its current standing in relation to the social constraints and the actions of others and itself.

A node stores the information produced by its ORM (such as its permissions and obligations) in a social state repository, which is part of the 'state & attributes' module. A node's *character* is also present, permitting higher-level behavioural traits to be specified, such as those giving a node's broad social outlook - e.g. law-abiding or anarchic, altruistic or reciprocal. This is similar to the agent-type and conflict resolution strategies of the BOID framework (Broersen et al., 2001, 2002) (see also the *Related Work* section). Pragmatic factors affecting a node's ability to participate in a network, such as communications range and battery power, are recorded in the physical attributes component, while other information relating to a node's reasoning process (e.g. perceptions) are stored in the intentional state component. The *physical attributes* state component captures AHN node aspects such as position, communications range and battery power; pragmatic factors affecting a node's ability to participate in a network.

The control module houses a protocol database of common interaction patterns (for example, *contract net* (Smith and Davis, 1978; Pitt et al., 2001)). These are used to process incoming messages and formulate appropriate responses, as well as to initiate conversations. A corresponding communications interface serves as point of dispatch and receipt of the actual messages. The originators and recipients of messages may be other agents, the environment or, in a controlling/monitoring capacity, the agent owner/user/experimenter.

At the time of writing, the outlined implementation has not yet been fully realised. Our current implementation proposal is based on a single *Prolog* process in which the interactions of multiple agents/nodes are simulated. Apart from the communications, there are several AHN characteristics that are of particular interest to us. These are *range* (physical constraints on node communications ability), *adaptability* (ability to cope with fluctuations in node participation) and *heterogeneity* (variation in physical and logical attributes of AHN nodes). As suggested previously, these aspects are to be captured by the *physical attributes*, ORM and *character* components respectively.

The current system architecture features a representation of the physical environment alongside multiple agent instances. The agents perform physical actions (such as 'moving') by making appropriate requests of the environment. The outcome of such attempts are determined by the environment's assessment of the corresponding action's viability. A request to move, for example, may only be honoured if (amongst other things) the intended destination is unoccupied.

A processing loop handles, in turn, the simulation cycles for the environment and other agents. The 'state' of individual agents is made available in the appropriate cycle, whereupon updates can take place. The decision-making procedure of each agent likewise occurs in the cycle. This includes consultation of the ORM. We do not, at this stage, provide details or a specific characterisation of the agent interpreter, as this has yet to be finalised. We nevertheless describe the ORM as a contributor to the deliberative process of each agent, calculating, based on the agent's most recent actions and observations, consequent and relevant normative positions. We note that rather than inherently constraining deliberation (in the way that the `generate_options` stage of a BDI/PRS implementation might), the ORM output is advisory: other factors (such as *character*) will contribute to the process.

Given the nature of the simulation model, it is possible to share a single ORM 'instance' between agents, as the narrational perspective of each agent is uniquely established within each cycle. The performance of the ORM in this respect may be augmented by the retention of some normative state aspects between cycles.

Inter-agent communication in the above system is modelled using a simple 'mail-box' approach. Messages are 'sent' by copying the content to queues (*Prolog* lists), again accessible within the corresponding program cycle. The `send_msg` and `recv_msg` messaging primitives are used to insert and remove messages from the relevant FIFO queue. This communications framework is intentionally generic so as to facilitate future migration to a different system model — for example, a multi-threaded or distributed one.

We are currently investigating how to express 'norms' within the proposed implementation. There are several existing frameworks that address similar architectural concerns (see, for example, (Castelfranchi et al., 1999; Stratulat et al., 2001; Boella and Damiano, 2002; Kollingbaum and Norman, 2003b)). We identify the norm formulation of (Stratulat et al., 2001) in particular as being particularly comprehensive. Key elements of a norm within the approach include:

- *type*: (for example) an obligation, right or power;
- *subject*: a non-empty set of roles or agents to whom the norm applies;
- *object*: an action or a state (property) referred to in the norm context;
- *authority*: the institutional context in which the norm exists;
- *author*: the 'issuer' of the norm (can differ from the *authority*);
- *interval of validity*: a time-period over which a norm is applicable;
- *context of application*: a logical pre-condition for a norm's application;
- *cost of violation*: a measurement of expected punishment for norm-violation;

Our initial approach will draw from the above feature-set, as deemed appropriate, in its representation of norms as *Prolog* terms. The normative positions appearing in Section 3 — such as $\mathsf{pow}(\mathsf{agentB}, \mathsf{act3})$ — can be expanded to include references to additional attributes as required. In their existing form, they can be thought of as a type of shorthand — for example, it may not be necessary to include *authority* as a norm parameter if it is deemed a constant of a particular simulation scenario.

We have not provided further details of the decision-making aspects of the agent architecture because these have yet to be finalised. In addition, the emphasis of our investigation lies in the incorporation of a norm-governed approach to AHN, rather than an agent-oriented approach. It is possible that we can validate the efficacy of a norm-governed approach in this respect without the need for a sophisticated agent model.

We propose the use of an external observer, as described earlier, to detect whether agents operate in compliance with their norms. In the outlined implementation, is is therefore required that the communications of all participating agents are appropriately monitored. This can be achieved through the introduction of a privileged entity (the society monitor) that is informed of all messages exchanged between the other agents. The observer can be implemented so that participating agents forward it copies of all sent messages, or it can form part of the communications substrate of the system, so that use of the messaging primitives implicitly informs the observer. In this way, the possibility exists of checking actions for compliance with the governing social constraints and acting on detected transgressions accordingly.

6 Protocols and policies for AHN management

We have identified three central processes facilitating AHN operation: *network formation*, *association* and *resource access control*. Network formation concerns the interactions occurring in forming, modifying and dissolving an AHN. Association refers to the establishment of task-specific groups within an AHN (effectively, a form of sub-grouping). Resource access control refers to the management and usage of resources within an association. In our norm-governed AHN extension, we identify corresponding protocol types for each issue process, namely *admission*, *session* and *task-oriented* (See Figure 5).

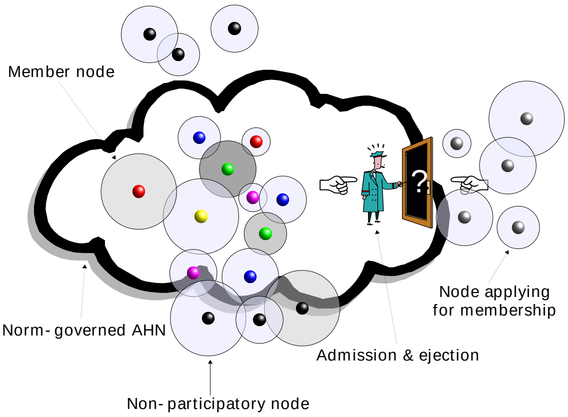


Figure 5: A norm-governed ad hoc network

An admission protocol allows nodes to create, enter and exit AHNs. In addition to establishing base communications requirements for interacting with the rest of the network, this protocol defines 'terms and conditions' that will govern such interaction. The behaviour of participants can be broadly regulated in this manner. The protocol likewise gives the joining node member basic assurances and expectations about the AHN. Participation in an AHN may come about through active or passive means. In the former case, a potential member initiates the admission protocol. In the latter case, an existing member of the AHN initiates proceedings by contacting a potential recruit. The protocol also provides basic tracking and migration services. Due to transmission and mobility issues, a participating node may come and go out of transmission range, intentionally or otherwise. In both cases, appropriate arrangements must be made, including resource reallocation, placeholding, address-forwarding, message storage, role re-allocation and traffic re-routing.

Session management protocols are used within an AHN to form groupings of nodes working on a common task (the dashed lines between nodes in Figure 5). Session management covers a broad range of concepts including resource discovery, session participation, quality of service monitoring and location tracking. For the most part, a session control protocol can be thought of as a more exclusive version of an admission protocol in that it achieves similar objectives but on a smaller and more exacting scale. For this reason, it is possible for the role of the latter to be subsumed by the former.

Each session may be subject to further specific task-oriented protocols, such as auction, contract net and floor control. It is possible for a node (agent) to participate in more than one session, assuming the respective roles held in each session are compatible. By a similar argument, an agent may be able to operate simultaneously in more than one society or AHN.

The roles occupied by agents in a session may not be fixed. Just as resources can be allocated in different ways, role assignment can be done through a dynamic procedure — such as an election. In addition, a role may be duplicated or shared among different agents in an AHN. For example, it might be that any existing member node is able to function as an 'admissions clerk' to any node wishing to join the network.

We have described each of the three protocols above in their generic forms. Different AHNs and agent sub-groupings may desire or require variants of these protocols. We consequently identify a *voting protocol* that can be used at the meta-level. Agents can use this protocol to decide on the specifics of any procedures used at a lower level. In this way, protocols can be tailored to local conventions and requirements. In addition, the voting protocol can itself undergo similar modification. This is to change the manner in which a vote takes place — choosing *quorum*, *mandatory* or *majority* voting constraints, for example.

Figure 6 illustrates the main relationships between the described protocols. The admission, session and task-oriented protocols are placed top-down, in the expected order of use and precedence. The role and voting protocols run alongside, adding further 'dimension'. This signifies the concept of allowing the roles that arise during execution of the other protocols to be (re-)assigned. The presence of the voting protocol as a final layer reflects the ability of subjects to collectively and procedurally determine how and when the rules governing interaction should change.

We have assumed a protocol-oriented approach for agent communication. While such a stance has clear precedence, we note that protocols alone are insufficient for specifying behaviour within an open agent society (Artikis, 2003). There is always a possibility of agents failing to comply, either through accident or for selfish reasons. This is especially the

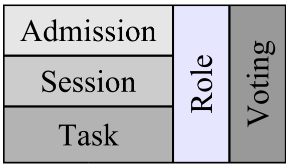


Figure 6: Norm-governed AHN protocol hierarchy

case in commercial and other competitive scenarios, where individual rather than collective needs are paramount. It is for this reason that we advocate a norm-governed approach to supplement the existing protocol-oriented one (Artikis et al., 2002; Pitt, 2003; Artikis et al., 2004).

7 Related work

Our norm-governed approach to AHNs proposal can be essentially characterised as an application of a normative framework to network management. There are, therefore, identifiable themes shared between this and other work within the same context. In this section, we conduct a representative (that is, non-exhaustive) review of formal frameworks, applied theory and applications that reflect or relate to the issues raised by our proposal.

7.1 Programmable and intelligent networks

This work takes place within the EPSRC Programmable Networks Initiative (*ProgNets*) and is therefore related to other network management research. A central ProgNets theme is that of *Active Networks* (Tennenhouse and Wetherall, 1996), which extend the conventional network model from that of solely data transport to include code execution. User programs can then be run on network nodes to shape data transport and delivery. Example technologies based upon this extended model include the *Safety Net* programming language (Wakeman et al., June 2001) and *cognitive packet networking* (Gelenbe et al., 2001). These have most impact at the physical network level. *MagnetOS* (Barr et al., 2002) addresses application-level concerns by introducing an AHN operating system layer that presents an unified and transparent application programming interface of the distributed AHN.

7.2 An agent society dynamics framework

Fasli provides a framework for formal analysis of social agent behaviour, in which the term ‘social agent’ refers to a group consisting of one or more agents (Fasli, 2003a,b). Within the framework, the normative concepts of roles, commitments, obligations and rights are introduced to regulate Multi-Agent Systems (MAS). In particular, commitments are positioned as a means to secure group cohesion while roles and social commitments are adopted in order to achieve individual objectives and collective agent commitments respectively. The framework builds upon a typical BDI logic, adding sorts for agents, social agents and domain objects. These provide the basis for formalising social agent behaviour. For example, Fasli interprets norms and rules as ‘general obligations that ought to be believed by everyone’.

There are also directed obligations — held between bearer and counterparty — with associated rights. Agents can be characterised according to whether they always seek recompense if possible (‘strict’) or if they refrain from immediate punitive action (‘lenient’), should the opportunity arise. The framework extends to role-related concepts, with each role having associated commitments, obligations and rights. Upon fulfilment of those commitments, the role can be disbanded.

The framework does not cater for heterogeneous agents with potentially competing agendas since the external behaviour specifications assume agent sincerity. This contrasts with the principles of OASs regarding the inability to determine individual agent motivation and internal state. The framework therefore seems more suited to co-operative scenarios where a certain degree of agent transparency is assumed.

7.3 The BOID architecture

Broersen *et al.* have proposed the Beliefs-Obligations-Intentions-Desires (BOID) architecture, with associated mechanisms to resolve conflicts between these mental attitudes (Broersen et al., 2001, 2002). Obligations and desires are respectively treated as external and internal motivational attitudes within the BOID framework. BOID’s architects identify the problem of deciding which obligations and desires to fulfil given an agent’s current belief and intentions as being especially important in a complex environment. The suggested solution is a conflict resolution strategy that prioritises updates according to the agent’s character. For example, when a clash occurs between a prior intention (that is, an as yet unfulfilled commitment to achieve some state of affairs) and a (current) belief, a *realistic* agent will abandon the intention: the agent believes it is now impossible to achieve the intention. Similarly, a *social* agent will prioritise the preservation and fulfilment of obligations (commitments to other agents) over its own desires. A number of resolution strategies are available; each is labelled by a permutation of the letters in 'BOID' to convey the relative weighting given to the mental attitudes under the particular scheme.

The agent's mental state is updated by an iterative, resource-bounded process that operates as follows. In the simplest case, the first step involves finding which, of set of rules mapping one proposition logic formula to another, are applicable ('triggered') by the current state. The rules are ranked by a function ρ that assigns a unique numeric score to each rule. Furthermore, ρ is strictly sorted with respect to mental attitudes: the scores for all rules concerning a specific attitude are either all less than or greater than the rule-scores for any other attitude. Thus ρ determines the agent type by returning scores for the precedence of each rule. Candidate rules are chosen based on score and consistency. Rule application yields a new *extension* that can serve as input to the same procedure; this continues until the agent runs out of resources (mainly time) or if it reaches a fixed point (rule closure). It is also possible to broaden the scope of the extension calculation procedure and modify ρ so as to allow more general agent types to be expressed within the framework.

The BOID framework's internal conflict resolution strategy is effective to the extent that it prevents any norm inconsistencies from ever arising. In certain circumstances, however, such inconsistencies may be tolerated. An agent may have norm inconsistencies from the adoption of conflicting roles, but in practice, be able to operate without problem.

7.4 The Normative Agent System model

In Stratulat et al. (2001), a first-order *normative agent system* (NAS) model is proposed for representing the performance of sequenced agent actions together with durative, dynamic norms. In NAS, norms are identified as a means to co-ordinate agent activities and include obligations, permissions and interdictions (prohibitions). Norms have a life-time and are used to influence (when prescribed by an authority such as a system architect or indeed another agent) and monitor (when the authority checks for compliance) agent behaviour. In the former case, agents are made aware of norms, but are at liberty to comply or ignore them. In this sense, NAS is applicable to OASs in that agents are externally influenced by norms independent of their internal construction. From the agent perspective, the motivation for norm-adherence is a quantifiable benefit arising from compliance and a corresponding, measurable punishment resulting from the discovery of norm violation.

The NAS framework incorporates the deontic concepts of obligation, permission and prohibition. Norms are expressed as conditional propositions, with a deontic formula as consequent and a contextual, first-order term as antecedent. The framework uses a time-model influenced by both Situation Calculus (Allen and Ferguson, 1994) and Event Calculus, with augmented syntax for agent actions. Norm-violation is detected by a model-checking technique that considers a history of actions and events together with a description of ideal system behaviour.

7.5 NoA

The Normative Architecture (NoA) framework (Kollingbaum and Norman, 2003a,b) equates norms with obligations, prohibitions and permissions. Within the language, obligations generate goals (an objective being to achieve a state of affairs) and actions (an objective being that action's performance). As duals of obligations, prohibitions generate objectives *not* to achieve a state of affairs. Finally, permissions allow for specific goal and action objectives to be pursued and fulfilled.

NoA introduces the concept of 'Norm Activation', whereby each normative statement has an activation condition (from when it applies) and expiration condition (from when it ceases to apply). Each statement features both these concepts together with a role specification (identifying the subject agent to whom the norm applies) and an activity specification. For obligations and permissions, the norm activation parameters respectively define when an obligation should be fulfilled and when precisely a permitted course of action can be undertaken. The authors identify norms as being 'active' or 'passive' depending on whether or not they currently apply. If a permission is currently passive, it does not mean that the associated action or state of affairs is 'forbidden'. Likewise, a deactivated prohibition does not mean that the corresponding action/state is 'allowed'.

The NoA architecture permits conflicting norms to co-exist and identifies three different types of (in)consistency in relation to the adoption of new norms and the actions and states of affairs that are the subjects of an agent's current obligations and prohibitions. These are: *strong inconsistency*, when the agent's revised normative state is inevitably inconsistent; *strong consistency*, when the revised norm-set is inevitably consistent; and *weak inconsistency*, when the revised norm-set is possibly inconsistent, but avoidably so. These classifications are used by an agent to determine the best course of action, by way of a *look-ahead* strategy that calculates the effect of specific norm adoption on the level of (in)consistency. Kollingbaum and Norman note that this strategy is computationally intensive as the temporal properties and sequencing of norm adoption/fulfilment, in addition to the exogenous events of a dynamic environment, must be taken into account. They identify trade-offs between language expressiveness and interpretation that can lead to more tractable applications (ignoring or restricting the temporal aspects of norms, for example).

The NoA framework is the only one of the frameworks reviewed here that addresses the concept of institutionalised power (Jones and Sergot, 1996). We believe that this is essential for modelling agent activities in a norm-governed system, requiring equal attention as given to other normative concepts such as permission and obligation.

7.6 Policy languages and other concepts

There are also other relevant network resource management technologies. The *Ponder* language (Damianou et al., 2001), for example, is used to specify security policies that can be interpreted to provide management strategy for computer networks, software and hardware. It incorporates features such as *authorisation*, *delegation*, *information filtering* and *refrain* to specify access control policy for a variety of resources in an object-oriented framework. Organisational aspects of the language include roles and management structures. Roles allow the grouping of policies involving a common subject while management structures package roles and the relationships between them. (Firozabadi et al., 2004b) also considers management structures, but from a *privilege management* perspective, whereby authority certificates are used for access control and delegation. A significant detail in this context is that of interoperation between distinct AHN, such as when an agent is a member of more than one network. (Firozabadi et al., 2004a) models resource-sharing between heterogeneous enterprises by way of coalition policy. They define a framework for governing resource-sharing between coalition members based on a policy language of obligation and entitlement.

We observe that the above specification and management approaches represent a strong case for the use of normative concepts in network and resource regulation. While the approaches use common terminology (such as obligation, permission and role), however, they individually make use of very different concepts. Terms used in an implementation-oriented approach, in particular, do not correspond to those used in a more formal approach.

8 Summary and further work

Ad hoc networks are communications structures with non-fixed topology operating without the need for the traditional infrastructural requirements associated with fixed network systems. AHNs are formed by the deliberate interaction of devices operating through a shared medium and protocols. Based on these characteristics, we argue that an AHN can be viewed as a form of open agent society. We seek to determine whether this analogy can be usefully exploited through the development and application of analytical tools and processes previously used in the OAS context to that of AHNs.

We propose to develop a logically sound, and computationally grounded, theoretical framework for institutional management of, and self-organisation in, ad hoc networks. In particular this demands the formal representation of a variety of normative concepts (Sergot, 2001; Pacheco and Carmo, 2003; Boella and der Torre, 2004) — in the first instance, permissions, obligations and institutional powers (Jones and Sergot, 1996) but also contractual and quasi-contractual duties and rights, arrogation and abrogation of power, appointment of surrogates, and delegation mechanisms. Each node in the ad hoc network is aware of its individual normative relations vis-à-vis other components, and uses this to manage its position within the network and its contribution to the overall application.

We plan to extend the existing formalisms and implementations of multiple, interacting, social and/or organisational structures and cultures to support additional normative concepts such as right, duty and mandate, and the arrogation and abrogation of power. We also plan to develop enhanced and/or alternative description languages to represent the richer relations required for ad hoc networks, and the associated methods of automated verification (such as model checking).

We also aim to address the engineering requirements that would ensure the above techniques scale up for pervasive computing applications (Kamara et al., 2004) methodological and tool support, logical and physical models for run-time knowledge distribution, platform and component architectures for direct manipulation of normative relations, and complex system dynamics concerning the changing network structure and normative relations over time. We plan to build a proof-of-concept demonstrator, facilitating systematic experimentation and evaluation of our framework.

Acknowledgements

We would like to thank the reviewers and Alexander Artikis for their feedback on earlier versions of this paper.

This research is is being carried out within the NoGoSoN project and is funded by EPSRC UK (GR/S69252/01) under the Programmable Networks Initiative.

References

James F. Allen and George Ferguson. Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4(5):531–579, 1994.

Alexander Artikis. *Executable Specification of Open Norm-Governed Computational Systems*. PhD thesis, University of London, 2003. URL `http://www.doc.ic.ac.uk/~aartikis/publications/artikis-phd.pdf`.

Alexander Artikis, Lloyd Kamara, Jeremy Pitt, and Marek Sergot. A protocol for resource sharing in norm-governed ad hoc networks. In *To appear in the Proceedings of the Declarative Agent Languages and Technologies (DALT) workshop, AAMAS'04*, 2004.

Alexander Artikis, Jeremy Pitt, and Marek Sergot. Animated specifications of computational societies. In Cristiano Castelfranchi and W. Lewis Johnson, editors, *AAMAS'02*, pages 1053–1062. ACM Press, 2002.

Alexander Artikis, Marek Sergot, and Jeremy Pitt. An executable specification of an argumentation protocol. In *Proceedings of the 9th International Conference on Artificial Intelligence and Law, ICAIL 2003*, 2003.

Rimon Barr, John C. Bicket, Daniel S. Dantas, Bowei Du, T. W. Danny Kim, Bing Zhou, and Emin Gün Sirer. On the need for system-level support for ad hoc and sensor networks. *SIGOPS OS Review*, 36 (2):1–5, 2002. ISSN 0163-5980.

Guido Boella and Rossana Damiano. An architecture for normative reactive agents. In *Proceedings of PRIMA 2002*, number 2413 in LNAI, pages 1–17. Springer-Verlag, 2002.

Guido Boella and L. Van der Torre. Regulative and constitutive norms in normative multiagent systems. In *KR'04*, 2004.

Jan Broersen, Mehdi Dastani, Joris Hulstijn, and Leon Van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3–4): 428–447, 2002.

Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leon Van der Torre. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *ICMAS'01*, pages 9–16. ACM Press, 2001. ISBN 1-58113-326-X.

Cristiano Castelfranchi, Frank Dignum, Catholijn Jonker, and Jan Treur. Deliberate normative agents: Principles and architecture. In *Proceedings of The Sixth International Workshop on Agent Theories, Architectures, and Languages (ATAL-99)*, Orlando, FL, 1999.

Nicodemos Damianou, Naranker Dulay, Emil Lupu, and Morris Sloman. The ponder policy specification language. In *Policy 2001: Workshop on Policies for Distributed Systems and Networks*, number 1995 in LNCS, pages 18–39. Springer-Verlag, 2001.

Maria Fasli. From social agents to multi-agent systems: Preliminary report. In Vladimír Marík, Jörg Müller, and Michal Pechoucek, editors, *Multi-Agent Systems and Applications III: 3rd International Central and Eastern European Conference on Multi-Agent Systems*, volume 2691 of *LNCS*, pages 111–121. Springer-Verlag, 2003a. ISBN 3-540-40450-3.

Maria Fasli. Towards an organisational approach to teamwork. In *Proceedings of the AAAI-02 Workshop on Autonomy, Delegation, and Control: From Inter-agent to Groups*, pages 988–989. ACM Press, 2003b. ISBN 1-58113-683-8.

Babak Sadighi Firozabadi, Marek Sergot, and Olav L. Bandmann. Using authority certificates to create management structures. In Bruce Christianson, Bruno Crispo, James A. Malcolm, and Michael

Roe, editors, *Security Protocols, 10th International Workshop, Cambridge, UK*, volume 2845 of *LNCS*, pages 134–145. Springer, 2004a. ISBN 3-540-20830-5.

Babak Sadighi Firozabadi, Anna Squicciarini, Marek Sergot, and Elisa Bertino. A framework for contractual resource sharing in coalitions. In *Fifth IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY'04)*, pages 117–126. IEEE, 2004b.

Erol Gelenbe, Ricardo Lent, and Zhiguang Xu. Design and analysis of cognitive packet networks. *Performance Evaluation*, pages 155–76, 2001.

Enrico Giunchiglia, Joohyung Lee, Norman McCain, Vladimir Lifschitz, and Hudson Turner. Nonmonotonic causal theories. *Artificial Intelligence*, 153 (1–2):49–104, 2003.

Andrew Jones and Marek Sergot. On the characterization of law and computer systems: The normative systems perspective. In J.-J. C. Meyer and R. J. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 275–307. Wiley, 1993.

Andrew Jones and Marek Sergot. A formal characterisation of institutionalised power. *Journal of the IGPL*, 4(3):429–445, 1996.

Lloyd Kamara, Alexander Artikis, Brendan Neville, and Jeremy Pitt. Simulating computational societies. In *Engineering Societies in the Agents World III: Third International Workshop Proceedings*, volume 2577 of *LNCS*, pages 53–67, 2003.

Lloyd Kamara, Jeremy Pitt, and Marek Sergot. Norm-aware agents for ad hoc networks: A position paper. In *Proceedings of the Ubiquitous Agents Workshop, AAMAS'04*, 2004.

Martin Kollingbaum and Timothy Norman. NoA — a normative agent architecture. In *IJCAI-03*, pages 1465–1466, 2003a.

Martin Kollingbaum and Timothy Norman. Norm adoption in the NoA agent architecture. In *AAMAS'03*, pages 1038–1039. ACM Press, 2003b. ISBN 1-58113-683-8.

Robert Kowalski and Marek Sergot. A logic-based calculus of events. *New Generation Computing*, 4 (1):67–95, 1986. ISSN 0288-3635.

Olga Pacheco and José Carmo. A role based model for the normative specification of organized collective agency and agents interaction. *AAMAS*, 6(2): 145–184, 2003. ISSN 1387-2532.

Charles E. Perkins. *Ad Hoc Networking*. Addison Wesley Professional, 2001.

Jeremy Pitt. Constitutive rules for agent communication languages. In Georg Gottlob and Toby Walsh, editors, *IJCAI-03*, pages 691–698. Morgan Kaufmann Publishers, 2003.

Jeremy Pitt, Lloyd Kamara, and Alexander Artikis. Interaction patterns and observable commitments in a multi-agent trading scenario. In *Proceedings of the fifth international conference on Autonomous agents*, pages 481–488. ACM Press, 2001. ISBN 1-58113-326-X.

John Searle. What is a speech act? In *The Philosophy of Language*, pages 39–53. Oxford University Press, 1971.

Marek Sergot. A computational theory of normative positions. *ACM Transactions on Computational Logic*, 2(4):581–622, 2001. ISSN 1529-3785.

Reid G. Smith and Randall Davis. Distributed problem solving: The contract-net approach. In *Proceedings of the Second Conference of Canadian Society for Computational Studies of Intelligence*, 1978.

Tiberiu Stratulat, Françoise Clérin-Debart, and Patrice Enjalbert. Norms and time in agent-based systems. In *Proceedings of the 8th international conference on Artificial intelligence and law*, pages 178–185. ACM Press, 2001. ISBN 1-58113-368-5.

David L. Tennenhouse and David J. Wetherall. Towards an active network architecture. *SIGCOMM Comput. Commun. Rev.*, 26(2):5–17, 1996. ISSN 0146-4833.

Ian Wakeman, Alan Jeffrey, Tim Owen, and Damyan Pepper. Safetynet: A language-based approach to programmable networks. *Computer Networks*, 36 (1):101–114, June 2001.

Jie Wu and Ivan Stojmenovic. Ad hoc networks. *IEEE Computer*, 18(9162):29–31, 2004.