Autonomous norm acceptance

Autonomous Norm Acceptance

Rosaria Conte [1], Cristiano Castelfranchi [1], Frank Dignum[2]

[1]Division of AI, Cognitive and Interaction Modelling - IP/Cnr - Rome, Italy - email: rosaria@pscs2.irmkant.rm.cnr.it

[2] Eindhoven University of Technology - The Netherlands - e-mail: dignum@win.tue.nl

Abstract. It is generally acknowledged that norms and normative action emphasize autonomy on the side of *decision*. But what about the autonomous *formation* of normative goals? This paper is intended to contribute to a theory of how agents form normative beliefs and goals, and to formulate general but *non* exhaustive principles of norm based autonomous agent-hood – namely goal generation and decision making- upon which to construct software agents.

1 Norms as Inputs to Goals

It is generally acknowledged that norms and normative action emphasize autonomy on the side of *decision*. But what about the autonomous *formation* of normative goals?

In a recent paper (Dignum & Conte 1997), the treatment of goal acquisition in the Agent Theory (AT) literature was found inadequate, some formal rules for goal generation have been proposed, and the role of social inputs in the acquisition of new goals has been emphasized. Here, we intend to continue that work, by including norms among the social inputs to one's goals, and by extending the goal generation rule to the case of normative goals. The general question then is, how and why do autonomous agents form normative goals? The answer to this question goes back to a former paper by some of the authors (Conte & Castelfranchi 1995), where a typology of reasons for accepting norms has been explored in analogy with goal adoption. Here, however, the formal instruments worked out with regard to the general rules for goal generation will be applied to the special case of normative goals. In the next section we will describe other work related to norms in the multi-agent field. In Section 3 the objectives of this work will be described. In Section 4, the main concepts necessary for the description of norm acceptance will be introduced, and in Section 5 the rules for goal formation will be summarized to serve as a basis for the

J.P. Müller et al. (Eds.): ATAL'98, LNAI 1555, pp. 99-112, 1999.

rules for norm acceptance. Finally, in Section 6, we will give a (formal) treatment of norm acceptance in which several aspects of autonomy in norm acceptance will be distinguished and characterized. Section 7 will conclude and indicate areas for further research.

2 The Main Issues

Many mechanisms have been proposed to regulate social relationships between agents in multi-agent systems. The interesting new concept of emotion was introduced in another paper in these proceedings (Bazzan *et al*. 1998). Through simulations of societies of agents in which some were altruistic and others egoistic they showed that altruistic behavior is beneficial for the whole society. Unfortunately, they did not indicate how altruistic behavior could evolve. Their agents were either altruistic or egoistic. Both emotions were hardwired into the behavior of the agents. In this paper we actually want to show how self-interested agents change their social behavior under the influence of their experiences (through the use of norms and norm violations).

Legal and social norms also have received a considerable attention in the social sciences, in logical and social philosophy, in some AI sub-fields (legal expert systems, norm-based reasoners, etc.), and more recently in the MAS domain. Nonetheless, they have not yet received a satisfactory explanation.

As for the social sciences, no theory of autonomous normative decision as grounded upon agents' internal representations has been provided. Norms are often viewed as *emergent* properties of utilitarian agents' behavior, independent of their beliefs and goals (Binmore 1994).

As for the logical models of obligations, the connections between obligations and mental states are usually not formalized (Shoham & Cousins 1994).

In the Multi-Agent Systems field, social norms are perceived to help improve coordination and cooperation (Shoham & Tennenholtz 1992; Jennings and Mamdani 1992; Conte & Castelfranchi 1995; Walker & Wooldridge 1995). Furthermore, the advent of large communication networks, civic networks, as well as the spread of electronic commerce, contributed dramatically to draw the attention of the AI scientific community to issues such as *authorization*, *access* regulation, *privacy* maintenance, respect of *decency*, etc. not to mention the more obvious problems associated with the regulation of the *use* and *purposes* of networks.

Indeed, the efforts done by MAS researchers and designers to construct *autonomous* agents (Wooldrige & Jennings 1995) carry with themselves a number of interesting but difficult tasks:

1. how to avoid interference and collisions (also metaphorical) among agents autonomously acting in a common space?
2. How to ensure that negotiations and transactions fulfil the norm of reciprocity? Imagine a software assistant delegated to conduct transactions on behalf of its user. In principle, due to its loyalty (benevolence), the assistant will behave as a shark

with regard to potential partners, always looking for the transaction most convenient for its user, and thereby infringing existing commitments.

3. More generally, how to obtain robust performance in team work (Cohen & Levesque 1990)? How to prevent agents from dropping their commitments, or better, how to prevent agents from disrupting the common activity (cf. Jennings 1994; Kinny & Georgeff 1994; Tambe 1996; Singh 1995)?

These questions have become central research issues within the MAS field. Other problems are perhaps less obvious. For example, the existence of so-called virtual *representatives* brings about the question of delegation. Software assistants, mobile agents are intended to act as virtual representatives of network clients. But the role of representatives implies that some normative mechanism is at work, such as *responsibility* (Jennings 1995) and *delegation* (Santos & Carmo 1996). Analogously, the concept of role (Werner 1990) and role-tasks - which is so crucial for the implementation of organizational work - requires a model of *authorization* and (institutional) *empowerment* (Jones & Sergot 1995).

These concepts explicitly or implicitly point to at least two questions, of vital importance:

1. how do agents acquire norms? In the formal social scientific field[1], the spread of norms and other cooperative behaviors is usually not explained by means of models of internal representations of norms. The object of inquiry usually consists of the conditions under which agents converge on behaviors which prove efficient in solving problems of coordination (Lewis 1969) or cooperation (Axelrod 1987). In the multi-agent field, norms are represented in the agents, but they are treated as built-in constraints. Therefore, what about the acquisition of new norms? This question is crucial with regard to all the problems listed above. If agents are enabled to acquire new norms, there is no need for expanding exceedingly the individual agents' knowledge base. Consequently, the multi-agent system may be optimized when it is *on-line*, while multi-agent systems where norms have been actually implemented allow for a modification of norms only when the system is *off-line* (Shoham & Tennenholz 1992).
2. How can agents violate norms? So far, norms are treated as constraints to either the agent's action repertoire (Shoham & Tenneholz 1992) or its evaluation module (Boman 1996). Norms operate by reducing the set of available or convenient actions to those that meet the existing constraints. Therefore, norms apply unfailingly. Agents cannot violate them. However, the possibility to violate norms is crucial for solving possible conflicts of norms, which often arise among tasks associated with different roles, or among norms belonging to different domains of

[1] That is, in utility theory and in game theory. Social (psychological) theorists have attempted behavioral explanations of normative influence. However, these theories cannot be immediately translated into computational models of autonomous norm-acceptance, since poor attention is paid within behavioral social science to the internal representations and processing of norms. On the other hand, cognitive social psychologists pay attention to rules of reasoning (natural vs. formal logics) rather than to moral and social norms. Generally speaking, the role of cognition for social action is still relatively poorly explored.

activity. This question is crucial with regard to both legal expert systems and autonomous agents interacting in a common world.

Both questions bring into play autonomy: the capacity for acquiring and the capacity for violating them are direct consequences of the agents' autonomy and bear crucial application consequences. If we need autonomous agents, we also need autonomous normative agents. One advantage of autonomous agents is their capacity to select the external requests, which it is necessary or convenient for them to fulfil. This selective capacity affects not only the agents' normative *decisions*, but also their *acquiring* new norms. Indeed, agents take a decision even when they decide to form a „normative belief", and then to form a new (normative) goal, and not only when they decide whether to execute it or not. The decision to form a normative goal will be called here norm *acceptance;* the decision to execute a norm will be called norm compliance. Obviously, this depend on a radical divorce of goals from intentions (see again Dignum & Conte 1997). Although we will not provide examples of implementation in this paper, computational applications of autonomous norm compliance do exist (think of the systems in which norms are treated as inputs for decision making; for a reference to these systems, see Boman, 1996). As was observed by Shoham and Tennenholtz (Shoham & Tennenholtz 1992), computational models of autonomous norm acceptance are lacking in the field of multi-agent systems. In our view, a capacity for autonomous norm acceptance would greatly enhance multi-agent systems' flexibility and dynamic potentials. To implement such a capacity it is necessary to model and implement the recognition of norms and the formation of normative beliefs.

Here, we will primarily deal with autonomous formation of new *normative* beliefs and goals. To do so, the more general property of *social* autonomy must be characterized.

3 Objectives

This paper is not intended to provide a descriptive theory of human agents' normative behavior. Instead, we intend to

1. contribute to a theory of autonomous normative decision as grounded upon agents' internal representations, and
2. formulate general but *non* exhaustive principles of norm based autonomous agent-hood, namely goal generation and decision making. These principles should be seen as applicable to both natural and artificial systems. Whether they are necessary and sufficient to describe the behavior of real natural systems is of no concern here. We aim at identifying mechanisms such that, if implemented into some artificial systems, will give rise to autonomous norm acceptance and compliance.
3. A sub-goal is to design principles for how software agents should be constructed in order to exhibit autonomous normative action. Empirical claims about the behavior of software agents that are designed according to these principles will be specified.

However, the empirical control of the validity of the present model is beyond the scope of this paper. Possible experimental controls through computer simulation are under study.

4 Concepts for Dealing with Normative Agents

In this section we will introduce the concepts necessary to define norm acceptance by agents.

An *agent* is a system whose behavior is neither *accidental* nor strictly *causal*, but oriented to achieve a given state of the world.

Goal-governed agents are able to achieve goals by themselves, by planning, executing, adapting and correcting actions. A goal-governed or purposive behavior (Miller et al. 1960; Rosenblueth & Wiener 1968) is controlled by goals. Agents that contain explicit representations for goals, intentions and beliefs are called *cognitive* agents.

Goals are internal explicit representations of world states which agents want to be realized.

Beliefs are those propositions about the world that agents hold to be true. In the rest of this paper we assume our agents to be cognitive agents.

Intentions are those goals that agents intend to reach, and intentional actions are those actions that agents intend to perform (Castelfranchi, 1995). Cognitive agents are not necessarily *autonomous*. Autonomy requires autonomous goals (Covrigaru & Lindsay 1991). It is a relational concept: a system is defined as autonomous always with regard to another system. An agent is autonomous only relative to other agents in a common world: *x is autonomous from y as for a goal p* where *p* is a behavior of *x* (for a cognitive agent, *p* is a goal). Here, we will consider only social autonomy, that is to say, *autonomy from other agents.* To be noted, autonomy is not a none-or-all notion. There are different levels and kinds of autonomy. With goal-governed agents, the most important distinction is relative to their level of autonomy. Here we will focus on goal autonomy and norm autonomy.

An agent x is *goal-autonomous* if and only if whatever new goal q it comes to adopt, there is at least a goal p of x's to which q is believed by that agent to be instrumental. More precisely, a socially autonomous agent adopts other agents' goals only if this adoption is conceived of as a way to achieve one or more further goals. As shown in (Dignum & Conte 1997), to adopt a goal does not imply to generate the relative intention and perform the relative action. It is also possible that an agent *adopts a given goal* but will not eventually pursue it; this does not become an intention, because, for example, *it is not preferred* to other more important goals.

A *norm* is an obligation on a given set of agents to accomplish (active norm), or abstain (passive norm) from, a given action. A norm is only external, when its subject agents have no mental representation, neither goal nor belief that corresponds to it. To be noted, norms are not meant here in the restrictive sense of laws, but in the more general sense of social obligations and conventions. Phenomena such as that of

reciprocity imply obligations and permissions even though they may not allow for a strictly juridical treatment.

4.1 Empirical Criteria for Autonomous Normative Agents

An agent is *norm-autonomous* if it can:
1. recognize a norm as a norm (normative belief formation);
2. argue whether a given norm does or does not concern its case; decide to accept the norm or not;
3. decide to comply with it or not (obey or violate);
4. take the initiative of re-issuing (prescribing) the norm, monitoring, evaluating and sanctioning the others' behavior relatively to the norm.

In this paper we will examine the main aspects of norm-autonomous agents. Whenever we use the word agent, we will therefore mean norm-autonomous agent.

5 Previous Work: Goal Generation and the Role of Social Inputs

In this section, we will summarize the work done in (Dignum & Conte 1997) in which a formal model was developed for goal formation. In the next section we will apply the model there developed to the case of norms.

The general intuitive idea on goal formation is that an agent might form a goal *p if* it already has a goal *q* and achieving *p* is in some way *instrumental* to achieving *q*. We say that p is instrumental for q, denoted by INSTR(p,q), if achieving p contributes to achieving q. This notion of instrumentality can be seen as a generalization of the idea of sub-goals. In the next section we will say something more about different types of instrumentality in the context of normative goals.

The general goal generation rule is formalized as follows:

$GOAL_X(q|r \;\; BEL_X(INSTR(p,q)) \quad C\text{-}GOAL_X(p|GOAL_X(q|r) \;\; r) \quad (1)$

I.e. if an agent x has a goal q as long as r is true and it believes that p is instrumental to achieving q then agent x has a candidate goal p as long as it has the goal q and r is true. If x's beliefs about the instrumentality are given and do not change, the above rule is completely endogenous. I.e. it does not depend on any external situation or change of circumstances. However, the goal generation rules can also be used to react to the environment. To effect this, the agent should have some beliefs about the benefits of reacting to other agents. I.e. how the generation of a goal in response to an event contributes to some overall goal of itself. In (Dignum & Conte 1997) three possible behaviors were given as input for goal formation:
1. behavioral conformity
2. goal conformity
3. goal adoption

Behavioral conformity is effected through the following two formulas:

$GOAL_X(\text{be-like}(x,y)|\text{true}) \quad (2)$

$BEL_X[DONE(y, \;) \quad INSTR(DONE(x, \;), \text{be\_like}(x,y))] \quad (3)$

It is easy to see that, with the goal generation rule, we can derive that x will do whatever y does as long as x wants to be like y. Or formally:

$C\text{-}GOAL_x(DONE(x,)|GOAL_x(be\_like(x,y)))$ (4)

The idea of goal conformity is similar to that of behavioral conformity, except that x will now mimic the goals of y. This is formalized by the following:

$GOAL_x(be\text{-}like(x,y)|true)$ (5)

$BEL_x[GOAL_y(p|r) \quad r \quad INSTR(p,be\_like(x,y))]$ (6)

And again with the goal generation rule we can derive:

$C\text{-}GOAL_x(p|GOAL_x(be\_like(x,y)))$ (7)

The idea of goal adoption is slightly different from the previous two. In this case, x not only takes over a goal, but it also tries to help y to obtain its goal. The formulas to describe this are as follows:

$GOAL_x(help(x,y)|true)$ (8)

$BEL_x[GOAL_y(p|r) \quad r \quad INSTR(help(x,y),OBT_y(p))]$ (9)

Therefore, whenever x believes that y has a goal p it will try to help y to obtain p.

As can be seen from the above, all types of goal formation follow the same pattern. Given some overall goal of x's, x believes that in some circumstances (y has performed some action or has some goal) it is instrumental for x to have some (candidate) goal that helps achieve the overall one. The (candidate) goal that will be generated depends on the type of behavioral rules that the agent follows.

In (Dignum & Conte 1997) a sketch of the semantics of the logic that is used above is given. Due to space limitations we will leave such formalization out of the present paper. In the next section we will explore whether similar rules as were given for goal formation can be used for autonomous norm acceptance.

6 Normative Inputs to One's Goals

Norms are an important device for some agents to influence and control the behaviors of other social agents, and thereby make the whole social behavior more predictable. In order to influence the behavior of the agent, a norm itself must generate a corresponding intention; and in order to generate an intention it must be adopted by the agent, and become one of its goals. First, the agent x must be aware that the norm is in force (belief) and concerns (belief) the agent itself; secondly, x must have some motive of its own to obey the norm, since in general x must have reasons for adopting goals from outside. Which are the motives for norm acceptance? What kind of autonomy is brought about by norm acceptance and by normative agents?

6.1 Forms of Autonomy in Norm Acceptance

There are two decisions to be taken in the process from a normative input to a conforming normative behavior (norm compliance): *the acceptance of the norm as a norm*; and *the decision to conform to it.*

Norm Recognition as Presupposition of Norm Acceptance

The issue is whether the agent will accept the candidate norm *as a norm*, and why it will accept it. For the purpose of this paper we will take the candidate norms to be external norms that are somehow observed by the agent but in reality several things can operate as candidate norms.

We will denote candidate norms as obligations: $O_{yX}(q)$, where q stands for the norm, y is the authority that issues the norm and X is the set of intended addressees of the norm (the norm subjects). An autonomous agent is able to evaluate a candidate norm against several criteria. It can reject it for several reasons:

1. *evaluation of the candidate norm;* if it is based upon [2] an already recognized norm, the norm is recognized as a norm itself ; if not
2. *evaluation of the source;* if the norm is not based upon a recognized norm, the entity y that has issued the norm is evaluated. If y is perceived to be entitled to issue some norms (it is a normative authority), $O_{yX}(q)$ can be accepted as a norm; this belief entails or is supported by other more specific beliefs relative to several of y's features:
   - (i) q is (not) within y's domain of normative competence;
   - (ii) the current context is (not) the proper context in which y is entitled to issue q ;
   - (iii) y is addressing a set of agents that is (not) within the scope of its authority.
3. *evaluation of the motives*; $O_{yX}(q)$ is issued for y's personal/private interest, rather than for the interest y is held to protect: if x believes that y's prescription is only due to some private desire, etc. x will not take it as a norm. x might ignore what the norm is for, what its utility is for the group or its institutions, but may expect that the norm is aimed at having a positive influence for the group; at least, it is necessary that x does not have the opposite belief, that is, that the norm is not aimed to be „good for" the group at large, but only for y. This is so crucial of a norm that one could even conceive it as implied by the first belief: y is entitled only to deliver prescriptions and permissions that are aimed at the general rather than at its own private interest.

The agent subject to $O_{yX}(q)$ is an *evaluator* of $O_{yX}(q)$. The output of its evaluation is a normative belief: the belief about the existence of a norm[3] (rather than of a simple request or expectation). We can formalize the evaluation process with the following two formulas:

(a) $BEL_x(O_{ZU}(r))$ $BEL_x(O_{ZU}(r)$ $O_{yX}(q))$ (10)

(b-c) $(O_{yX}(q)$ $BEL_x(auth(y,X,q,C))$ $BEL_x(mot(y,OK)))$
$BEL_x(O_{yX}(q))$ (11)

Both formulas lead to $BEL_x(O_{yX}(q))$. The first through simple modus ponens and the second directly from its fulfilled premises. Many things can be said about when one norm implies another (see e.g. (Royakkers 1996 and Herrestad & Krogh 1996)),

[2] The new norm is just an instantiation, application, or interpretation of the former one.

[3] Notice that such an evaluation and recognition plays a very active role as one step of the process of collective norms *creation*: to recognize that a given norm exists as a norm puts it into existence (see later).

but to go into this subject is beyond the scope of this paper. The relation "auth" introduced above stands for: y is authorized to issue the norm q to the set of agents X in context C. The relation "mot" indicates that the motives of y are indeed correct. Both relations are of course very complex. More about the authorization can be found in (Dignum & Weigand 1995).

The acceptance of the norm as a norm is an act that contributes both to spreading around the norm in question as well as to constructing/creating/forming the norm at the social level.

Norm Acceptance

Once a norm has been recognized as a norm, a normative belief has been formed. x has an additional belief. Is such a belief sufficient for the formation of a new goal? The answer to this question that we can derive from our postulate of social autonomy is: No! A normative belief is never sufficient for the formation of a new goal. Another ingredient is needed, that is, a goal already formed in x's mind for which x believes that complying with the norm n is instrumental.

Social autonomy has a normative corollary: *A norm-autonomous agent accepts a norm q, only if it sees accepting q as a way of achieving (one of) its own further goal(s).*

$$BEL_X(O_{yX}(q) \quad INSTR(OBT_X(q),p) \quad GOAL_X(p|r)) \rightarrow N\text{-}GOAL_X(OBT_X(q)|GOAL_X(p|r) \quad r) \quad (12)$$

Intuitively, the above formula states that x forms a normative goal $OBT_X(q)$ (i.e. accepts the norm q) if x believes that the norm exists (for agents in set X) and that fulfilling the norm (i.e. $OBT_X(q)$) is instrumental to one of its own goals. Although the rule for norm acceptance resembles the one for goal formation there are a few important differences. The first difference with the goal formation rule is that in the premises we included a belief of an existing norm. I.e. a normative goal is only derived with this rule if there exists some norm outside the agent to start with. Note that the implication in the rule is only a one-way implication. This means that not every normative goal has to be derived through this rule! We can imagine that agents can also autonomously form new norms. We could describe this by saying that the agent believes that a certain norm should exist, which leads to the following rule:

$$BEL_X(O(O_{yX}(q)) \quad INSTR(OBT_X(q),p) \quad GOAL_X(p|r)) \rightarrow N\text{-}GOAL_X(OBT_X(q)|GOAL_X(p|r) \quad r) \quad (13)$$

Where O(x) stands for a general obligation that holds for the set of all agents for which the issuer is a standard (central) authority. However, this is only one possible way in which new norms can be formed. We leave further discussion of this topic for another paper.

The other, less conspicuous, difference with the goal formation rule is the fact that we do not require q to be instrumental for the goal p, but rather $OBT_X(q)$. With $OBT_X(q)$ in this context we mean the fulfillment of the norm q by all members of X. The difference is that in this case we only try to fulfil the norm, because it is a norm. A much stronger case is when the norm itself is believed to act to the benefit of our goal p. This corresponds to "internalizing" the norm, making it our own goal, and would formally be described by:

$BEL_x(O_{yX}(q) \quad INSTR(q,p) \quad GOAL_x(p|r))$
$C\text{-}GOAL_x(q|GOAL_x(p|r) \quad r)$ (14)

This follows directly from the goal formation rule, because we have only strengthened the antecedent by adding a normative belief.

Given the above rule(s) for norm acceptance, it seems reasonable to see whether there are similar rules for norm conformity and norm adoption as were defined for goals. Obviously, we cannot define the same type of rules for norms, because an independent belief in the existence of some external norm is required before a normative goal is derived. x cannot deduce the existence of a norm by y performing a given action. Therefore we need at least the following two implications:

$BEL_x(BEL_y(O_{zX}(q))) \quad BEL_x(O_{zX}(q))$ (15)

$BEL_x(N\text{-}GOAL_y(OBT_X(q)|r) \quad INSTR(OBT_X(q),be\_like(x,y)))$ (16)

plus of course:

$GOAL_x(be\_like(x,y)|true)$ (17)

From the above, only norm adoption but no norm conformity is derived. It is not possible to mimic only the norms that were accepted by another agent! They should also be accepted in some way. Therefore we do have norm adoption, but no conformity. Of course, we can have "apparent" norm adoption in case an agent x adopts all the goals of an agent y that follow from a certain norm. In that case, if agent y fulfils the norm then agent x will follow it and fulfil the norm!

6.2 Norm Compliance

Once accepted, a norm becomes a normative goal. We distinguish normative goals from candidate goals primarily because the agent has different motivations to either choose a candidate or a normative goal as the goal it will actually try to achieve. The decision of normative compliance is influenced by the type of instrumentality of the norm, which is always related to some external source (the external norm). The candidate goals have an instrumentality that is determined by the inner motives of the agent. This difference becomes clear if we look at the reasons to give up a goal. If it is a normative goal it can be dropped at the moment the norm is changed or is no longer applicable. Candidate goals are only dropped when the agent knows they can no longer be achieved or a more urgent goal has become active.

Below, some reasons for non-conforming behaviors are summarized based on the different instrumentality evaluations described in the previous section:

1. *Norm responsibility;:* the agent has accepted the norm $O_{yX}(q)$, but is only prepared to try to fulfil this norm itself. It will not try to "help" other agents to fulfil the norm. Formally:
2. $N\text{-}GOAL_x(OBT_X(q)|r) \quad C\text{-}GOAL_x(OBT_x(q)|r)$ (18).
3. *Goal conflict*: the normative goal contrasts with goals that are more urgent than the goal of complying with the norm. The expected value of norm violation depends on factors that vary with different kinds of agents, societies, or situations; such factors include

1. the importance of the goal or *value* of respecting the norms, of being a good citizen, etc.;
2. the importance of possible *feelings* related to norm violation (guilt, indignity, etc.);
3. the importance of foreseen *negative consequences of the violation for the global interest* that the norm claims to protect, or for other important societal goals (e.g., to violate norms will destroy respect, trust, and solidarity in the society).
4. the probability and weight of *punishment* (including social approval and its consequences);

5. *Norm conflict* (ubi major...): these may include provocation and rebellion, or other normative goals prescribing opposite norms (e.g., pacifist vs. military norms).
6. *Impertinence*: x does not believe to be a member of the set of agents mentioned by the norm; for example, x strongly supports the norms regulating the car traffic, but has no driving license. Obviously, x can be said to execute the norm at a higher level: it will probably support the norms in question by monitoring the drivers' behaviors any time it happens to have the possibility to do so. However, x will not execute the norm on its own.
7. *Material impossibility:* obviously when the norm prescribes an action which cannot be executed, x will not comply with it although it has recognized it as a norm and no conflict holds between the norm and other goals of x's; consider, for example, the case in which x finds itself trapped in a traffic jam. The traffic light turns red while x is in the middle of the crossing. x knows that it is violating the norm; it has recognized the norm, and has accepted it; x may even have formed a corresponding intention. Still, its behavior does not, and cannot correspond to what the norm prescribes.
8. If x accepts and executes a norm, it will monitor and check that people (subject to the same norm) respect it, and will implicitly or explicitly prescribe this, probably reacting to any violation of it, since this also turns into a frustration of a goal and expectation of x's.

Therefore, acceptance contributes to the *spreading* of the norm. Indeed norm spreading:

1. is not primarily behavioral but mentalistic: norms spread among minds through recognition (normative beliefs), acceptance (normative goals) and possibly, through *norm sharing* (see below);
2. the mental spread of norms will determine conforming behaviors which will influence the others and enhance the general acceptance and conformity.

Unlike current social scientific theories of norms, the hypothesis presented in this paper states that agents converge on norms if, and only if, these norms spread through their minds, and that there is a continuous feedback from some agents' norm compliance to the observation of this behavior by other agents, their forming new normative beliefs and goals, and their possible consequent norm compliance. While in current social scientific theories, the mental processing of norms is essentially overlooked for the emergence and diffusion of conventions, here it is considered as a crucial segment in the process leading to the emergence and spread of norms as a specific social cognitive artifact.

7 Concluding Remarks and Computational Applications

Here, we have endeavored to account for a process of autonomous normative decision, which includes two fundamental steps: the formation of a normative belief, and the decision to accept a norm. Indeed, not only to comply with a norm, but also to believe that something *is* a norm, are outputs of a complex decision making of an autonomous agent.

But the analysis described so far shows that a lot is yet to be done. In particular, two further aspects, mentioned throughout the paper, seem to play a fundamental role in norm spreading and emergence, especially in the case of social norms:

1. norm sharing: in some circumstances, agents come to share for some reason the utility, convenience or functionality of the norms. Under which circumstances does norm sharing occur? What are its effects?
2. autonomous norm formation: norms come into existence not only when they are "issued" by some „legislator“, but also when they emerge from agents' implicit or explicit agreements. How is this possible? In which moment does a social (non-institutional) norm start to exist and why?

The comprehension of both phenomena would largely benefit from the view presented in this paper: norms cannot be shared, without a mental representation of them. Analogously, while habits, routines, etc. emerge from a mere behavioral convergence, norms an only come into existence if agents start to believe that some given behaviors are obligatory and legitimate, that is, normative. The question is, how and why do agents form such beliefs. Both norm sharing and norm formation will be addressed in future studies.

Three types of computational applications of the model presented in this paper are under study:

1. in the area of legal expert systems, the implementation of (aspects of) the present model would allow for detection of
   - violations, which will be distinguished from other behaviors that do not correspond to norms; this would be of special interest in applications of expert systems to legal "diagnostic" reasoning and certification;
   - norm conflict; this is particularly relevant for the automatic advice to legal interpretation.
2. In the MAS field, the implementation of rules of norm acceptance would allow for
   - the acquisition of norms when the system is on-line, and consequently
   - greater flexibility of the system
   - a reduced load at the level of the agents' knowledge bases.
3. In the area of computer simulation, the hypothesis formulated in this paper can be experimentally tested; in particular, the advantages of autonomous norm acceptance and compliance in agents' monitoring others' behaviors should be tested and compared with models of convergence among autonomous but non-normative agents.

Acknowledgements

We would like to thank Harko Verhagen for convincing us about the opportunity and importance to make the level of autonomy postulated in our theory of norm- and goal adoption more explicit.

References

Axelrod, R. (1987). The Evolution of Strategies in the Iterated Prisoner's Dilemma. In L.D. Davis (ed) *Genetic Algorithms and simulated annealing*. Los Altos, CA: Kaufmann, 32-41.

Bazzan, A.L.C., Bordini, R.H., Campbell, J.A. (1998). Moral sentiments in multi-agent systems. In this volume.

Binmore, K. (1994). *Game-theory and Social Contract*. Vol. 1: Fair Playing. Cambridge: Clarendon.

Boman, M. (1996). Implementing Norms through Normative Advice, in R. Conte & R. Falcone (eds) *ICMAS '96 WS5 on "Norms, Obligations, and Conventions"*, Kyoto, Keihanna Plaza 10 Dec. 1996.

Cohen, Ph. & Levesque, H. (1990). Intention is Choice with Commitment. *Artificial Intelligence*, 42(3), 213-261.

Conte, R. & Castelfranchi, C. (1995). *Cognitive and Social Action*. London: UCL Press.

Covrigaru, A. A. & Lindsay, R.K. (1991). Deterministic Autonomous Systems. *AI Magazine*, Fall, 110-17.

Dignum, F. & Conte, R. (1997). Intentional Agents and Goal Formation. In M.P. Singh et.al. (ed) *Proceedings of the 4th International Workshop on Agent Theories Architectures and Languages*, Providence, USA.

Dignum, F. & Weigand, H. (1995). Communication and Deontic Logic. In R. Wieringa & R. Feenstra (eds) *Information Systems, Correctness and Reusability*. Singapore: World Scientific, 242--260.

Herrestad, H. & Krogh, C. (1996). Deontic Logic Relativised to Bearers and Counterparties. In J. Bing & O. Torrund (eds), *Anniversary Anthology in Computers and Law*, , Tano A.S, 453-522.

Jennings N.R. (1995). Commitment and Conventions: the Foundation of Coordination in Multi-Agent Systems. *The Knowledge Engineering Review* , 8, 223-250.

Jennings, N.R. (1992). On Being Responsible, in Y. Demazeau & E. Werner (eds) *Decentralized Artificial Intelligence 3*, Amsterdam: Elsevier Science Publisher, 93-102.

Jennings, N.R. & Mamdani, E.H. (1992). Using Joint Responsibility to Coordinate Collaborative Problem Solving in Dynamic Environments. In *Proceedings of the 10th National Conference on Artificial Intelligence*, San Mateo, CA: Kaufmann, 269-275.

Jones, A.J.I. & Sergot, M. (1995). Norm-Governed and Institutionalised Agent Interaction, *Proceedings of ModelAge'95: general meeting of ESPRIT wg* 8319, Sophia Antipolis: France, January, 22-24.

Lewis, D. (1969).*Convention*. Cambridge, MA: Harvard University Press.

Kinny, D. & Georgeff, M. (1994). Commitment and Effectiveness of Situated Agents. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, IJCAI-93, Sydney, 82-88.

Miller, G., Galanter, E., & Pribram, K.H. (1960). *Plans and the Structure of Behavior*, New York: Holt, Rinehart & Winston.

Rao, A.S. & Georgeff, M.P. (1991). Modelling Rational Agents within a BDI Architecture. In J. Allen, R. Fikes, & E. Sandewall (eds), *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, San Mateo, CA: Kaufmann, 473-485.

Rosenblueth, A. & Wiener, N. (1968). Purposeful and Non-Purposeful Behavior. In W. Buckley (ed.) *Modern Systems Research for the Behavioral Scientist*. Chicago: Aldine.

Royakkers, L. (1996). *Representing Legal Rules in Deontic Logic*. Ph.D. Thesis, Tilburg University, The Netherlands.

Santos, F. & Carmo, J. (1996). Indirect Action, Influence and Responsibility, in Brown, M. & Carmo, J. (eds), *Deontic Logic, Agency and Normative Systems*, Berlin: Springer, 194-215.

Shoham, Y. & Cousins, S.B. (1994). Logics of Mental Attitudes in AI. In G. Lakemeyer & B. Nebel (eds) *Foundations of Knowledge Representation and Reasoning*, Berlin: Springer.

Shoham, Y. & Tennenholtz M. (1992). On the Synthesis of Useful Social Laws in Artificial Societies. *Proceedings of the 10th National Conference on Artificial Intelligence*, San Mateo, CA: Kaufmann, 276-282.

Singh, M.P. (1995). *Multi-Agent Systems: A Theoretical Framework for Intentions, Know-how, and Communications*. Berlin: Springer.

Tambe, M. (1996). Teamwork in Real-World, Dynamic Environments. In *Proceedings of ICMAS 1996*, Menlo Park, CA: AAAI.

Verhagen, H.J.E. & Smit, R.A. (1996). Modeling Social Agents in a Multi-Agent World, Eindhoven: Working Notes MAAMAW 1996.

Walker, A. & Wooldridge, M. (1995). Understanding the Emergence of Conventions in Multi-Agent Systems, *Proceedings of the First International Conference on Multi-Agent Systems*, the MIT Press, 384-389.

Werner, E. (1990). Cooperating Agents: A Unified Theory of Communication and Social Structure. In L.Gasser and M.N.Huhns (eds), Distributed Artificial Intelligence: Volume II. Kaufmann.

Wooldridge, M. & Jennings, N.R. (1995) (eds). *Intelligent Agents (LNAI Volume, 890).* Berlin: Springer.