Autonomous Norm-acceptance

Rosaria Conte Cristiano Castelfranchi Frank Dignum

Abstract

1 The acquisition of new goals: the case of norms

It is generally acknowledged that norms and normative action emphasise 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 emphasised. 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 main concepts necessary for the description of norm acceptance will be introduced. In section 4 the rules for goal formation are summarised to serve as a basis for the rules for norm acceptance. Finally, in section 5, we will give a (formal) treatment of norm acceptance in which several points of autonomy in norm acceptance will be distinguished and characterised. Section 6 will conclude and indicate areas for further research.

2 Current treatment of norms

The use of a normative vocabulary (obligations, deontic operators, norms, etc.) and the implementation of norms in expert systems have a wellestablished tradition in some AI sub-fields (legal expert systems, norm-based reasoners, etc.).

In adjacent domains (logical philosophy, social philosophy, decision theory), both legal and social norms have received a considerable attention. Nonetheless, this phenomenon has not received so far 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 either viewed as *emergent* properties of utilitarian agents' behaviour, independent of their beliefs and goals (Binmore, 1994), or as driven by a built-in moral sense. As for the logical models of obligations, the connections between obligations and mental states are usually not formalised (Shoham & Cousins 1994).

The advent of large communication networks, civic networks, as well as the spread of electronic commerce, contributed dramatically to draw the attention of the 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. These issues have a normative interpretation, since norms are not meant in the restrictive sense of laws, but in the more general sense of social obligations and conventions. Phenomena such as that of authorization imply obligations and permissions but do not allow for a juridical treatment.

In the Multi-Agent Systems field, social norms are perceived to help improve coordination and cooperation (Shoham & Tenneholz 1992; Jennings and Mandami 1992; Conte & Catselfranchi 1995; Jennings 1994; Walker & Wooldridge 1995). 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:

(a) how to avoid interferences and collisions (also metaphorical) among agents autonomously acting in a common space?

(b) 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. Due to its loyalty (benevolence), the assistant will behave as a shark with regard to potential partners, always looking for the most convenient transaction, and thereby infringing existing commitments.

(c) More generally, how to obtain a robust performance in teamworks (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 199?)?

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).

All these (logical-theoretical) models share two underlying questions, of vital importance:

1) how do agents acquire norms? In the formal social scientific field[1], the spread of norms and other cooperative behaviours is usually not explained by means of models of internal representations of norms. The object of inquiry usually the conditions for agents' convergence on behaviours which proved efficient in solving problems of coordination (Lewis 1969) or cooperation (Axelrod 1987). multiagent field, norms are treated as if they were built-in constraints. But 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 multiagent systems where norms have been actually implemented allow for a modification of the agents' norm-based representations only when the system is *off-line* (Shoham & Tenneholz 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). They operate by reducing the set of available or convenient actions to a subset of actions, that is, to those which 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 activity. Therefore, this question is crucial with regard to both leagl expert systems and autonomous agents interacting in a common world.

[1] That is, in utility theory and in game theory. Social (psychological) theorists have attempted behavioural explanations of normative influence (Homans 1974). However, these theories cannot be immediately translated into computational models of autonomous norm-acceptance, since poor attention is paid within behavioural social science to the internal representations and processing of norms. On the other hand, cognitive social psychology, the attention is focused on agents learning different rules of reasoning (natural vs. forma logics) rather than on moral and social norms. Generally speaking, the role of cognition for social action is still relatively poorly explored.

Both questions bring into play autonomy: the capacity for acquiring and violating norms are direct consequences of the agents’ autonomy and bear crucial applicative consequences. If we need autonomous agents, we also need autonomous normative agents. One advantage of autonomous agents is their capacity to filter external requests. Such a filtering capacity affects not only normative *decision*, but also the acquisition of new norms, which we will call here, norm-*acceptance*. The question is not only how agents take normative decisions, but also how they construct alternatives for such a decision. Violation, and for that matters obedience, not only depends upon autonomous decision among existing alternatives, but also from autonomous construction of alternatives. To state it differently, 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. 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). Conversely, computational models of autonomous norm-acceptance are lacking in the field of multiagent systems. As was observed by Shoham and Tenneholz (1992), a capacity for autonomous norm-acceptance would greatly enhance multiagent systems' flexibility and dynamic potentials. To implement such a capacity is conditioned to modelling and implementing agents' capacity to form normative beliefs, to recognize norms.

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

3 Objectives

This paper is not intended to provide a descriptive theory of human agents' normative behaviour. We intend to contribute to a theory of autonomous normative decision as grounded upon agents' internal representations, and formulate general but *not* exhaustive principles of norm-based autonomous agenthood, 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 behaviour 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.

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 behaviour of software agents that are designed according to these principles should 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. We will start with a definition of agents.

An *agent* is a system whose behaviour is neither *casual* 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 behaviour (Miller et al. 1960; Rosenblueth & Wiener 1968) is controlled by internal explicit representations (*Goals*). Agents that contain explicit representations for goals, intentions and beliefs are called *cognitive* agents.

Intentions are those goals that agents intend to reach, and intentional actions are those actions that agents intend to perform (Castelfranchi, 1995).

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

However, cognitive agents are not necessarily autonomous. Autonomy requires autonomous goals (Covrigaru & Lindsay 1991). It is a relational concept (Castelfranchi 1995): a system is defined as autonomus with relation to another system. Moreover, autonomys is a social concept. An agent is autonomous only relative to other agents in a common world: *x is autonomous from y as for her goal p* where *p* is a behaviour of *x* (for a cognitive agent, *p* is a goal). For example, a robot has some initiative, if it can refuse to do what the user aks it to do. In this case, its goals are autonomous from its user's; the robot in question has "its own" goals. 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.

4.1 Empirical criteria for autonomus normative agents

An agent is *norm-autonomous* if it can:

(a) recognise or not a norm as a norm (normative belief formation);
(b) argue whether a given norm concerns or not its case; decide to accept the norm or not;
(c) decide to comply or not with it (obey or violate);
(d) take the initiative of re-issuing (prescribing) the norm, monitoring, evaluating and sanctioning the others' behaviour 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 summarise 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 subgoals. In the next section we will say a bit more about types of instrumentality in the context of normative goals.

The general goal generation rule is formalized as follows:

$$\mathrm{GOAL}_X(q|r)\#\ \mathrm{BEL}_X(\mathrm{INSTR}(p,q))\ 2\ \mathrm{C\text{-}GOAL}_X(p|\mathrm{GOAL}_X(q|r)\ \#\ r) \qquad (1)$$

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 endogeneous. 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 behaviours were given as input for goal formation:

1. behavioural conformity
2. goal conformity
3. goal adoption

Behavioural conformity is effected through the following two formulas:

GOAL$_x$(be-like(x,y)|true) (2)

BEL$_x$[DONE(y,å) 2 INSTR(DONE(x,å), be_like(x,y))] (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-GOAL$_x$(DONE(x,å)|GOAL$_x$(be_like(x,y))) (4)

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

GOAL$_x$(be-like(x,y)|true) (5)

BEL$_x$[GOAL$_y$(p|r) # r 2 INSTR(p,be_like(x,y))] (6)

And again with the goal generation rule we can derive:

C-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) # r 2 INSTR(OBT$_y$(p),help(x,y))] (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 achieving the overall one. The (candidate) goal that will be generated depends on the type of behavioural rules 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 a 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.

5 Normative inputs to one's goals

Norms are an important device for some agents to influence and control the behaviours of other social agents, and thereby make the whole social behavior more predictable. In order to influence the behaviour 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?

5.1 Forms of autonomy in norm-acceptance

There are two decisions to be taken in the process from a normative input to a conforming normative behaviour (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. In reality several things can operate as candidate norms, but a theory about the decision on what might form a candidate norm is beyond the scope of this paper.

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:

(a) *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

(b) *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 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 ;

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

(iii) y is addressing a set of agents that is (not) within the scope of its authority.

(c)*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) $BEL_x(O_{zU}(\ r))\ \#\ BEL_x(O_{zU}(\ r)\ 2\ O_{yX}(\ q))$ (10)

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

$2\ BEL_x(O_{yX}(\ q))$ (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)), but to go into this subject id 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

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

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).*

$$BEL_x(O_{yX}(\ q)\#\ INSTR(OBT_X(q),p)\ \#\ GOAL_x(p|r))$$
$$2\ N\text{-}GOAL_x(OBT_X(q)|GOAL_x(p|r)\#\ r) \qquad (12)$$

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:

$$BEL_x(O(O_{yX}(\ q))\ \#\ INSTR(OBT_X(q),p)\ \#\ \ GOAL_x(p|r))$$
$$2\ N\text{-}GOAL_x(OBT_X(q)|GOAL_x(p|r)\#\ r) \qquad (13)$$

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 conspicious, 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 fulfilment 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. We could also have the much stronger case in which we believe that the norm itself is to the benefit of our goal p. This is somehow "internalising" the norm and making it our own goal. This would formally be described by:

$$BEL_x(O_{yX}(\ q)\ \#\ INSTR(q,p)\ \#\ \ GOAL_x(p|r))$$
$$2\ C\text{-}GOAL_x(q|GOAL_x(p|r)\ \#\ r) \qquad (14)$$

We see that this would also follow 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:

$$BEL_x(BEL_y(O_{zx}(\ q)))\ 2\ BEL_x(O_{zx}(\ q)) \qquad (15)$$

$$BEL_x(N\text{-}GOAL_y(OBT_x(q)|\ r)\ 2\ INSTR(OBT_x(q),be\_like(x,y))) \qquad (16)$$

plus of course:

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

From the above, we can see that we can only have some form of norm-adoption and not norm-conformity. 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 fulfills the norm then agent x will follow it and fulfill the norm!

A last question is, what are the possible means-end links between q and p? In the case of goals we have left the instrumentality relation very open. However, in the case of norms this relation depends very much on the nature of p. We are far from providing an exhaustive typology. However, let us distinguish at least the following categories

(a) **Arbitrated instrumentality**: x sees accepting q as a means to obtain a non-natural consequent reward. In particular, avoidance of punishment, social praise, etc..

(i)*avoidance of external punishment*. Motives belonging to such category vary along two intertwined dimensions

- centralized/personal vs distributed/impersonal: this dimension refers to the existence of a personal authorship, an identifiable source of normative control, be it one's parents, one's spouse, one's best friend, the priest, a policeman, etc.
- institutional vs informal. Obviously, the present dimension is very close to the former, although they do not perfectly overalp. Indeed, an identifiable source can be institutional (explicitly empowered to do the controlling) or not. One's best friend or idol, the group leader, etc. can act as controllers with no need for a specific investiture. On the other hand,

decentralised control can be, although such a case is quite rare, institutionally empowered: in Hawthorne's famous novel, the Community is explicitly required by the Council of Seniors to inflict the punishment to the adulterous everytime she appears in public settings. Usually, however, identifiable controllers are institutionally and deliberatley empowered to play such a role, while decentralized contol is spontaenous.

(ii) *Avoidance of internal punishment*. These two categories are placed on a continuum. There is no radical difference between them. One could say that the more informal and distributed the source of punishment, the more it tends to be internalized.

(iii) *Achievement of external reward*. This should not be seen only as the positive side of the first category: x may comply with norms before external controllers not only to avoid punishment but also to cut a fine figure, to be looked after by its followers, to win the political campaign and beat its opponents, to impress its well-bred girl-friend, etc..

(iv) *Achievement of internal reward*. Again, this leads not only to improve or restore one's self-image, but also please one's moral and aesthetic sense, etc..

(b) **Natural instrumentality**:

(i) *Self-interested*: x sees accepting the norm as a means to obtain a natural consequent benefit: for example, pedestrians may want drivers to respect pedestrian areas.

(ii) *Value-oriented*: Social order and coordination, global benefit, social or distributive justice, solidarity, etc.. In Italy, now, some people want that taxes be paid in order for the country's deficit to be reduced (and thereby the chances to enter the European single currency to increase). This reason for norm-acceptance may be also seen as based upon "norm-sharing".

5.2 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 a 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 behaviours are summarised based on the different instrumentality evaluations described in the previous section:

(a) *Norm-responsibility;* the agent has accepted the norm $O_{yX}(q)$, but is only prepared to try to fulfill this norm itself. It will not try to "help" other agents to fulfill the norm. Formally:

$$\text{N-GOAL}_X(\text{OBT}_X(q) | r)\ 2\ \text{C-GOAL}_X(\text{OBT}_X(q) | r) \tag{18}$$

(b) *Goal-conflict*: the normative goal contrasts with goals that are more urgent than the goal of complying with norm. The expected value of norm violation depends on factors that vary with different kinds of agents, societies, or situations; such factors include

> (i) the probability and weight of *punishment* (including social approval and its consequences);
>
> (ii) the importance of the goal or *value* of respecting the norms, of being a good citizen, etc.
>
> (iii) the importance of possible *feelings* related to norm violation (guilt, indignity, etc.)
>
> (iv) 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).

(c)*Norm-conflict* (ubi major...); these may include provocation and rebellion, or other normative goals prescribing opposite norms(e.g., pacifist vs military norms);

(d) *Unpertinence*: 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 licence. Obvioulsy, x can be said to execute the norm at a higher level: it will probably support the norms in question by monitoring the drivers' behaviours any time it happens to have the possibility to do so. However, x will not execute the norm on its own.

(e)*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 entrapped in a traffic jam. The trafficlight turns red while x is in the middle of the crossing. x knows that he is violating the norm; he has recognized the norm, and has accepted it; x may have even formed a corresponding intention. Still, its behaviour does not, and cannot correspond to what the norm prescribes.

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, which 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:

a) is not primarily behavioural but mentalistic: norms spread among minds through recognition (normative beliefs), acceptance (normative goals) and possibly, through *norm sharing* (see below);

b) the mental spread of norms will determine conforming behaviours which will influence the others and enhance the general acceptance and conformity.

6 Concluding remarks

Here, we have endeavoured 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: norm-sharing and autonomous norm-formation. Let us spend some words on both issues in turn.

Norm-sharing

There are different types of norm-sharing depending on the level at which the sharing occurs

(a) *sharing the means*: x sees q as a good solution to a given end but rejects that very end;

(b) *sharing the goal*: x may share what is perceived to be the end of the norm; (c) *sharing some meta-goal*: x may share q at some meta-level, e.g., trusting a specific authorship; reinforcing any authorship; trusting norms; reinforcing compliance with norms; sharing norms as such; etc..

These different types have different effects in the process of norm-acceptance, upon which one should investigate. Now, norm-sharing is by no means necessary for a norm to be accepted, let alone complied with. Why bother with it? Because, norm-sharing seems to be a sufficient (although not necessary) condition for social control. Norm-acceptance per se does not guarantee such a result, since x is unlikely to monitor others' obedience with norms before having chosen in favour of its own obedience. However, even if, say, I am not personally involved in the norms controlling safety while driving motorcycles (e.g., wearing helmets) them, I may share them and put pressure on youngsters to observe them.

Norm-formation

Without a general *recognition,* a social norm is not a norm. (At the legal level it is sufficient that the authority is recognised as authority and that the norm is recognised as correctly issued by it). Autonomous agents subject to norms are in fact autonomous norm creators. They create norms through their evaluation and recognition, through their compliance, and through their interpersonal issuing, monitoring and judging.

Norms are multi-facets and MA objects requiring different role players: *norm issuing* (the “legislator”); *norm acceptance or recognition* (the observer or the addressee); *norm obedience* or *compliance*, or violation (the addressee); *norm monitoring* (the “policemen”, the judge, the bystander); *normative punishment* or approval (the policemen, the bystanders). Now the creation or formation of the norm occurs not only when the “legislators” is issuing them. Social norms are collective cooperative constructions, based upon implicit or explicit agreement, convergence and expectations. Therefore, even norm recognition is in fact an act of norm formation, and this act is autonomous.

Consider also that the “legislator” is not necessarily an “official” institutional person. Its function may be a *decentralised role*: any agent that prescribes a given behaviour as a norm on other agents (or on itself), is in fact issuing (or re-issuing) that norm. Since any norm-addressee that accepts the norm (and in particular, that obeys it) wants the norm to be obeyed by the other addressees, each (respectful) addressee becomes also an informal legislator and an inspector (Conte & Castelfranchi 1995). Indeed, norm-formation is but a continuous, spontaneous, decentralised process executed by autonomous normative agents in interaction.

Acknowledgements

We would like to thank Harko Verhagen for convincing us -both in his paper (Verhagen & Smit 1996) and in a discussion - about the opportunity and importance to make the level of autonomy postulated in our theory of norm- and goal-adoption more explicit. Harko would like to see *conscious* agents explicitly represented, in order to capture this level of autonomous agency. We think that even without an explicit and formal theory of consciousness it is possible to capture high levels of autonomy and cognitive norm processing.

References

Boman, M. 1996. Implementing norms through normative advice, in R. Conte and R. Falcone 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, UCL Press, London.

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. Singh et.al., editor, Proceedings of the 4th International workshop on Agent Theories Architectures and Languages, Providence, USA, 1997.

Dignum, F. & Weigand, H. 1995. Communication and deontic logic. In R. Wieringa and R. Feenstra, editors, *Information Systems, Correctness and Reusability*, pages 242--260. World Scientific: Singapore.

Herrestad, H. & Krogh, C. 1996. Deontic Logic relativised to bearers and counterparties. In J. Bing and O. Torrund, eds, *Anniversary Anthology in Computers and Law*, pages 453-522, Tano A.S.

Jennings N. 1995. Commitment and Conventions: the foundation of coordination in multi-agent systems. The Knowledge Engineering Review , 8.

Jennings, N. 1992. On being responsible, in Decentralized Artificial Intelligence 3 (Elsevier Science Publisher, Amsterdam) 93-102.

Jennings, N. R. & Mandami, E. H. 1992. Using joint responsibility to coordinate collaborative problem solving in dynamic environments. In Proceedings of the 10th National Conference on Artificial Intelligence, 269-275, San Mateo, California: Kaufmann.

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.

Kinny, D. & Georgeff, M. 1991. Commitment and effectiveness of situated agents. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, IJCAI-93, 82-88, Sydney, .

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 Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning, J. Allen, R. Fikes, E. Sandewall (eds), 473-485, San Mateo, California: Kaufmann.

Rosenblueth, A. & Wiener, N. 1968. Purposeful and Non-Purposeful Behavior. In Modern systems research for the behavioral scientist, Buckley, W. (ed.). 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. and Carmo, J. (eds), Deontic Logic, Agency and Normative Systems, 194-215, 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, 276-282. San Mateo, California: Kaufmann.

Singh, M.P. 1995. Multi-agent Systems: A Theoretical Framework for Intentions, Know-how, and Communications. Springer Verlag, LNCS, volume 799.

Tambe, M. 1996. Teamwork in real-world, dynamic environments. in Proceedings of ICMAS 1996, Menlo Park, California: AAAI.

Verhagen, H.J.E. & Smit, R.A. 1996. Modelling social agents in a multiagent world, Working Notes MAAMAW 1996, Eindhoven.

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, editors, Distributed Artificial Intelligence: Volume II. Morgan Kaufmann Publishers.

Wooldridge, M. & Jennings, N. 1995 (eds). Intelligent Agents (LNAI Volume, 890). Springer-Verlag.