Artificial Intelligence and Law **7:** 323–340, 1999.
© 1999 *Kluwer Academic Publishers. Printed in the Netherlands.*

From conventions to prescriptions. Towards an integrated view of norms[*]

ROSARIA CONTE and CRISTIANO CASTELFRANCHI
PSCS-Project for the Simulation of Social Behaviour, Institute of Psychology, Cnr-National Research Council, V.le Marx 15 - 00137 Rome, Italy

Abstract. In this paper, a model of norms as cognitive objects is applied to establish connections between social conventions and prescriptions. Relevant literature on this issue, especially found in AI and the social sciences, will be shown to suffer from a dychotomic view: a conventionalistic view proposed by rationality and AI scientists; and a prescriptive view proposed by some philosophers of law (Kelsen 1934/1979, Hart 1961, Ross, 1958).

In the present work, the attempt is made to fill the gap between these views by putting forward a hypothesis concerning the process from perceived behavioural regularities to normative assumptions. The emergence of norms will be here seen as intrinsically intertwined with the emergence of normative beliefs. Unlike that assumed by the conventionalistic sight, the process of emergence is seen as a non-continuous phenomenon. A given behavioural regularity will be argued to give rise to a normative belief if and as long as that regularity is believed to be prescribed within the community. Two corollaries of this hypothesis will be examined: (1) unlike that implied by the conventionalistic view, the spreading of norms is not only due to a passive behavioural social influence (imitation) but also to an active cognitive one (the spreading of normative wants and beliefs); (2) unlike that assumed by the prescriptive view, a norm is not necessarily explicitly and deliberately issued by some normative authority, but is grounded upon the norm-addressees' beliefs that they are generally prescribed to comply with it.

> One day, a sentinel in charge with watching a besieged castle spread for fun a false alarm about a forthcoming enemy. But as he saw the population getting in arm and running to the city walls, he himself hastened to defend his city from the enemy he had invented.
> (Medieval tale)

Key words: agent, autonomy, belief, goal, prescription, reasoning

1. A Need for a Unified Theory of Norms

When approaching the issue of norms, their formal representation, and the explanation of their emergence, one is led to wonder how deliberate prescriptions (such

[*] This work has been realized with contribution of the ESPRIT Working Group on MODELling AGEncy.

as positive, explicit, legal norms) are related with social, implicit, customary norms or conventions.

Typically, the study of norms has focussed on either aspect, without providing a unification of these two phenomena. The emphasis on conventions prevails in the rational approach (Lewis 1969; Schelling 1960, etc.) and in the logics of obligations employed both in the Multi-Agent (MA) field (cf. Shoham & Tennenholz 1992, Jennings 1993) as well as in the area of Artificial Normative Reasoning (ANR) (cf. Brook 1994). Within the rational approach, let us distinguish two distinct views, which will be here called the epiphenomenal and the evolutionary views.

In the *epiphenomenal* view, norms are treated as convergence phenomena in populations of agents each oriented to maximize their utilities. In order to explain the emergence and spreading of conventions, game-theorists (see for example, Bicchieri 1990; for a review, cf. Kerr 1995), resort to a supposed tendency to conformity of social agents. In other words, within the rational approach, norms are seen as an epiphenomenon of both rational (read, self-interested) and conforming behaviour. While rationality is called up to explain the emergence of conventions, conformity is called up to explain their spread.

Authors (Schotter, 1981; Heiner, 1983; 1986) who have dealt with the *evolution* of social institutions, do not share such epiphenomenal view. As Heiner (1983) argues, rational decision theory and equilibrium models are not concerned with the evolution of social institutions, including norms. He claims that institutions (such as property rights, etc.) must evolve in order for agents to cope with uncertainty and limited information; they "*... enable each agent in the society to know less and less about the behavior of other agents and about the complex interdependencies generated by their interaction*" (p. 580; italics of the author). Here, social institutions are seen as useful tools socially evolved for coping with bounded rationality. Although certainly more elaborated than the preceding, such a view is based on a narrow conception of the role of cognition for social aims. Social institutions are seen as a result and a compensation for humans' limited mental capacities.

Within the MA field, and the ANR especially, instead, norms are treated neither as mere epiphenomena nor as means for compensating cognitive deficiencies, but as mental objects. In particular, in the work of Shoham and Tennenholz, a social norm is an operator of action, which, on given conditions, constrains the agents' action repertoire, that is to say, reducing them to a subset which is compatible with some socially useful state of affairs (e.g., avoid collisons). In this sense, the *representational* view of norms has a fundamental advantage over both the *epiphenomenal* and the *evolutionary*: it allows us to distinguish regular, customary behaviours from normative ones by identifying the cognitive correlates of norms, that is, to say by situating norms also in the minds of the social agents.

But, on the other hand, all these views concentrate on social norms, neglecting the deliberate issuing of positive norms. What is the relationship between these two phenomena, if any? Somehow, they have been considered essentially different, unrelated questions.

The fundamental *theoretical* consequence of such a dychotomy consisted in overlooking the prescriptive nature of norms. As will be shown thoroughout the paper, the *trait d'union* between legal and social norms consists in their common prescriptive nature and force. But this *trait d'union* is feasible only if prescriptions are distinguished from deliberately issued norms (otherwise, only legal norms can have a prescriptive force).

The most relevant consequence at the *formal* level consisted in leaving out the study of the interplay between norms and other mental operators. Even within the representational view, obligations have nothing to do with mental states. So far, there has been no clear connection between deontic operators and operators for mental states, such as epistemic operators (cf. Conte 1994, Conte & Castelfranchi 1995a, 1995b).

The consequence at the level of *design* consisted in building up *non-autonomous normative agents*. In Shoham and Tennenholz's work (1992), a normative agent is an artificial system that has an internal representation of norms, meant as socially useful laws. In this perspective, a normative agent is an agent which is endowed with the internal capacity to apply a (set of) norm(s). However, in the system developed by Shoham and Tennenholz, norms are implemented as built-in mechanisms. Therefore, agents have a sort of natural tendency to act in a socially useful way. They are not allowed to *decide* whether to comply with a norm or not. They can but apply norms. How is *transgression* possible, then? What is more, how is it possible for agents to *learn* a new norm? Only an autonomous normative agent is one which can acquire new norms, or violate the old ones. But existing artificial normative systems are not autonomous.

On the other hand, the view of norms worked out on the prescriptive side cannot be applied to customary norms. We will refer to Kelsen as a representative of this view (cf. Kelsen 1934/79). However, for the sake of argumentation, we will brutally simplfy his view, rendering it by far more extreme than it is in real matters.[1] Our purpose is to distinguish and illustrate two *implicit* conceptions of norms, rather than those *explicitly* stated within the relevant literature. Indeed, it should be acknowledged that attempts to integrate these two conceptions exist (Ross 1958, Hart 1961). For example, Hart (1961) had a far-reaching intuition when he stated that a fundamental bridge between the two views lies in the *internal* (read, mental) nature of norms. However, his intuition provided no real advance in filling the gap between conventions and prescriptions, simply because it could not profit from the theoretical and methodological tools of cognitive science. Only a *general* (read, multi-purpose and multi-level, applicable to human and non-human, natural

[1] Indeed, the theory expressed by Kelsen is much more complex and not fully consistent, since over his lifetime, this author has shifted from a rigidly prescriptive view to one much closer to the conventionalistic position. However, this author has provided the sharpest critique of the conventionalistic approach, and is therefore a good representative of the opposite view.

and artificial, individual and supra-individual minds), *constructive* (allowing to implement actual systems), and *integrated* (meaning, designing a consistent, comprehensive architecture of a working system, as required for that system to *act* in a concrete environment) model of social agency can fill the hole between conventions and prescriptions. In fact, the crucial bridge between prescriptions and conventions is offered by the systems' mental representations, and especially by their *goals*. But again, only the cognitive science can contribute to such an integrated model of social agency, which takes into account various motivations, interaction mechanisms and ingredients, coupling not only conformity and autonomy, but also imitation and influencing, coordination and cooperation, etc.. For example, an integrated model of social agency is necessary for modelling agents as likely not only to *imitate*, but also to *influence*, others. Therefore, it will investigate the spread not only of mere *regularities*, but also of social *control*. In other words, we propose to reverse the classical rational approach (be it epiphenomenal or evolutionary): rather than considering norms only as the effect of the agents' bounded rationality, and likelihood to interfere negatively with one another, we will consider the agents' cognition as a tool at the benefit of institutions, and particularly of norms.

Here, we will not wonder about what Kelsen nor other representatives of the prescriptive view have *really* said, but rather about the essence and implications of such a view. Greatly simplifying the issue, we can say that the bulk of the prescriptive view consists of asserting that norms necessarily imply a will, some goal. A norm is said to express the meaning of a concrete and specific act of volition, or to state *what someone wants someone else* (actually a set of people) to do. In the view proposed by Kelsen, there are no norms without such acts of volition: there are no norms without someone wanting, and therefore issuing, them.

Let us examine more carefully both views. As we shall see, both are unsatisfactory for opposite reasons. Later, we will turn to some attempts to integrate them.

1.1. A WEAK VIEW: NORMS AS CONVENTIONS

This view defines norms exclusively in terms of observable frequencies (or, more precisely, in terms of equilibria within a given population). It has most representatives in philosophy of rationality (Lewis 1969) and more specifically in game-theory (cf. Ullman-Margalit 1977; for a recent formulation, see Bicchieri 1990).[2]

The prescriptive view is correct when it says that norms are grounded upon some will, some goal. Simple matters of fact don't allow what is *normative* to be distinguished from what is *normal* (for the importance of such a distinction, see again Kelsen 1934/79, pp. 3 and 4, ch. 1). From the very fact that, in a given community, people *usually* throw their rubbish out of the window, one cannot draw the conclusion that they *ought* to do so, *even if things are such that the behaviour*

[2] However, in the philosophy of law, illustrious attempts to ground norms on customs and conventions abound (for one example, Ross 1958).

in question (throwing rubbish out of the window) *is a stable strategy*, one that is not likely to be invaded by different strategies. Moreover, one is not entitled to say that people *ought* to do so even if they *want* to conform to it. This conclusion is unwarranted for a number of reasons (for a conclusive analysis of them, see again Hart 1961):

(i) Intuitively, one knows that what *happens* is not a sufficient ground for deriving what *ought* to happen, whatever this means. You know that, however customary, throwing rubbish out of your window is not prescribed; actually, it is contrary to duty.

(ii) People *pretend* to understand the meaning of norms, their *raison d'être*, in order to comply with them. This may not actually be the case. Norms do not need to be cognitively shared in order to be respected. For a norm to be efficacious, it suffice to comply with it in order to avoid the expected sanctions. For example, one is not expected to share the norm of circulation when stopping at the red light. It is sufficient that one complies with it only to avoid taking a fine. Still, without a theoretical notion of normative as distinct from normal, one is unable to account for the above claim.

(iii) A behavioural strategy may be steady and resist invasion from other strategies when it is not *convenient* to unilaterally deviate from the strategy in question. If, to stick to our example, a person who refrains from throwing rubbish out of the window enters the population, s/he is bound to give in after a given time (because whatever she does everybody else will stick to their behaviors, which is an equilibrium; cf. Axelrod 1984, Bicchieri 1990) and conform to the majority. However, the calculus of one's convenience in choosing whether or not to *conform* to one given option has nothing to do with what is usually meant by *conformity*. The calculus of one's convenience does always occur in autonomous agents' decisions. The interesting question, at least from the point of view of a theory of norms, arises as to which options are there. Does a specific conformity come into play for its own sake, for the sake of complying with norms, or simply because that particular choice is convenient in terms of means/end reasoning? Obviously the last choice will always be more stable than any other option, since conformity may change under different social conditions, and compliance may yield to convenience. But this does not tell us that a choice beased upon convenience is also normative. Let us consider the following example: you are walking in the rain under your umbrella. Suddenly, you see people around you closing their umbrellas. Guessing that it stopped raining, you do the same. In such a case, your behaviour has been suggested by that of others, from which you inferred what is more convenient for you to do. It has nothing to do with norms. It is a rational behaviour, although it appears as a conforming behaviour.

As is well known, what *ought* to be the case cannot be derived from what *is* the case. However, two attenuating considerations are needed here. First, the process from regularities to norms, although non-continuous, is not even a none-or-all one. An analysis of the intermediate phenomena would be extremely interesting. For example, there are quite a number of social expectations (typical looks, manners, clothes, usually employed as identification markers especially among young people) which are empirically close to social norms, since they are in fact mandatory, but are not perceived as prescribed.

Secondly, it is to some extent true that the higher a given behavioural regularity, the more likely it will be prescribed. However, this observation should not be seen as an argument in favour of a conventionalistic view. Rather, it reminds us of the general social tendency to discourage violation of expectations, eccentricity, etc.. Indeed, on the grounds of such tendency, people assume that violations of expectations are strongly disapproved, that conformity is not only expected but also prescribed. Under this assumption, people will not conform out of pure conformism (if we are permitted the pun), but in order to comply with a social prescription. In other words, they will not simply imitate others' behaviours, but will accept their requests.

One ought to be fair towards game-theorists and acknowledge their attempt to ground social norms on firmer grounds than what is allowed by a view strictly based upon strategic equilibria. In the classical theory of convention (Lewis 1969), indeed, social norms are traced back to social problems (namely, problems of coordination). Now, this point deserves a careful consideration.

Within the classical conventionalistic view, norms are based upon some general goal,[3] or at least some distributive utility,[4] namely to achieve a better coordination among agents, and avoid interferences. Two questions arise here.

First, what is the very process leading to the assessment of a given conventional solution to the process of coordination? If a convention is a perceived solution to a *perceived* problem, it is based on some means/end reasoning. But where did that reasoning take place? Who came up with that solution? Saying that a convention is a *rational* solution to a problem of coordination does not equal to saying that the emergence of norms is a non-continuous process, where the cognitive mediation is essential?

Secondly, and more importantly, how to relate this view of conventions to the preceding one, which defines them in terms of stable strategies? Apparently, the link is quite clear. A stable strategy is one which each agent will be more likely to conform to. But a rational solution to a problem of coordination can be adopted by the agents either under the assumption that it is the most *convenient* solution to the problem, in which case there is no real conformity but a typical means/end reasoning. Or, it is adopted out of *imitation* of the others' behaviours, independently of any assumption relative to its utility. But in this case, we would not need any rational explanation: if a given strategy spreads *via* mere conforming attitudes, its virtual rationality is irrelevant for its spreading. Or, finally, it is adopted because

[3] One ought to examine the question of general, or even collective mental states. But this is not the forum for addressing such an issue. However, cf. Conte and Castelfranchi 1995a).

[4] This might look closer to the prescriptive view than it is. However, in the weak view, norms are reduced to, rather than derived from, distributed/collective utilities.

it is perceived as prescribed. Each answer renders the others superfluous. On the contrary, in the conventionalistic approach, the first two answers are mixed up. Of the two mechanisms considered, imitation and means/end reasoning, only the former can be regarded as responsible for the social *spreading* of conventions. Therefore, within a conventionalistic approach the utilitarian explanation is in fact redundant. It does not explain the adoption of the norm *and* its spreading. The model of spreading is too weak and non-predictive: a norm is what will prove to be a stable strategy.

1.2. THE ACT OF NORMATIVE PRODUCTION IS TOO MUCH

If the conventionalistic view is too weak, the prescriptive one is ambiguous and lends itself to a quite restrictive interpretation. Kelsen (1934/79), for example, proposed a rather limiting view of the will underlying a norm. In his terms, a norm always presupposes a volition issuing it, called *act of normative production*. From now on, this restricitve reading of the prescriptive view will be called *imperativistic*. Within such narrow confines, norms can only be intended as positive norms, while customary, social norms are ruled out.

However, our claim is that the imperativistic version of the prescriptive theory is unnecessary. This claim was made by other authors as well. To make but one example, Hart (1961) claims that commands become norms when they are "acknowledged" as such. In other words, norms do not necessarily imply an *act of volition*, but they are grounded upon an *act of recognition*. This claim is interesting but still insufficient. What indeed is an act of recognition? Moreover, what is the object of such a recognition?

A norm exists if it is required by some need or want. But this does not mean that this want turns into an imperativistic act of normative production, a deliberate and institutional issuing of the norm in question. If this consequence were necessary, we would be bound to rule out social norms, which are by definition spontaneous.

2. A Bridge between Conventions and Prescriptions

In the following, we will put forward the hypothesis that the dychotomous theory of norms discussed so far may be replaced with a unifying theory. In order to do so, a notion of norm must be found out allowing to throw a bridge between conventions and prescriptions, social and positive, implicit and explicit, customary and deliberately issued norms.

In turn, in order to realise this integration, several ingredients are required:

(i) a notion of goal sufficiently abstract and general as to be put in relation with obligations;

(ii) a notion of prescription which does not imply an explicit issuing, an act of normative production, and which is applicable to implicit norms, and therefore

(iii) a perspective complementary to that implied by the imperativistic view. The emergence of norms will not be seen from the point of view of a Legislator, issuing the norm, but from the point of view of the norm Addressee, deciding whether to comply with a given norm.

In the following, we will examine, in turn, (a) the variety of goals underlying a norm, and (b) the reasons why the current formal treatment of obligations does not deal with relating obligations with mental states. But beforehand, our formalism will be illustrated briefly.

2.1. OUR FORMALISM

The formalism used is a simplified version of Cohen and Levesque's (1990) language for describing their theory of rational action.

2.1.1. *Cohen & Levesque's Formal Theory*

One of the most influential theories of intentions, at least in the area of Multi-Agent Systems was developed by Cohen and Levesque (from now on, C&L) (1990).

This theory is aimed at modelling the "rational" properties of action. Intelligent, autonomous, rational agents are designed so as to be capable of producing and dropping intentions under *given* conditions. But which conditions are relevant for intentions formation and discharge? The authors developed an *incremental* view of intentions such that, at any step in a *goal*-driven process leading to intentions, agents are bound to decide, on the grounds of some relevant criterion, whether to keep to or abandon their goals. The language appears as a first-order language with operators for mental attitudes and action. They introduced two modalities for beliefs and goals,

(*BEL x p*) and (*GOAL x p*)

defined according to the possible worlds semantics, and therefore through accessibility relations. They implemented two modalities for action

(*(HAPPENS e*) and (*DONE a*)

expressing, respectively, events taking place in the world independent of the agents' actions and occurrence of actions. Finally, time is represented as an infinite sequence of events.

Beliefs and goals are given the usual possible world interpretation. As for consistency, the Hintikka axioms for beliefs apply to this model (see Halpern & Moses 1985). As for realism, goals are a subset of beliefs. (The accessibility relation ***G***, which defines the set of worlds in which goals are achieved is a subset of the accessibility relation ***B***, which defines the set of worlds belief-accessible to a given agent.). In such a model, in fact, a goal is defined as a belief-compatible desire. (In other words, agents cannot have goals which they believe to be unachievable.)

Many notions can be constructed on the grounds of these primitive modalities plus the operators $\Diamond$ for ‘later”, ; for “sequence” and ? for the procedure to test whether a given proposition is true.

$(\mathit{HAPPENS}\ a)$	an action will happen next
$(\mathit{DONE}\ a)$	an action has just happened;
$(\mathit{BEL}\ x\ \ p)$	x has p as a belief
$(\mathit{GOAL}\ x\ \ p)$	x has p as a goal;
$(\mathit{OUGHT}\ p)$	there is an obligation whatsoever on proposition p;
$(\mathit{AGT}\ x\ \ e)$	x is the only agent of the sequence e;
$e_1 \leq e_2$	e_1 occurs before e_2
p?	test action
$\Diamond p$	p will be true at some point in the future

2.1.2. *Molecular Predicates*

In the following, we present a number of definitions grounded upon the above atomic predicates that are necessary to understand the formulae provided throughout the paper. Most of them are drawn from Cohen and Levesque’s model, and we present them here for the convenience of the reader unacquainted with that model. Some have been introduced by the authors and other collaborators in preceding works (Conte et al. 1991; Castelfranchi et al. 1992).

$$(\mathit{DOES}\ x\ a) \overset{def}{=} (\mathit{HAPPENS}\ a) \wedge (\mathit{AGT}\ x\ a) \tag{1}$$

This says that x *is the only agent of action* a*, which will happen next.* We need an analogous predicate for past actions,

$$(\mathit{DONE\text{–}BY}\ x\ a) \overset{def}{=} (\mathit{DONE}\ a) \wedge (\mathit{AGT}\ x\ a) \tag{2}$$

saying that, x *is the only agent of action* a*, which has just happened.*

Cohen and Levesque have also introduced the following predicate to refer to sequences of world states,

$$\begin{aligned}(\mathit{BEFORE}\ q\ \ p)\ \ \overset{def}{=}\ \ & \forall c(\mathit{HAPPENS}\ c;\ p?) \supset \exists a(a \leq c)\\ & \wedge(\mathit{HAPPENS}\ a;\ q?)\end{aligned} \tag{3}$$

In words, q *comes before* p *when, for all events* c *after which* p *is true, there has been at least one event* a *preceding* c*, after which* q *was true.*

As for goals, Cohen and Levesque have introduced the notion of achievement goal, which is defined as follows:

$$(\mathit{A\text{–}GOAL}\ x\ \ p) \overset{def}{=} (\mathit{BEL}\ x\ \neg p) \wedge (\mathit{GOAL}\ x\ \Diamond p) \tag{4}$$

that is, *x has an achievement goal p if x believes that p is not true now but wants it to eventually become true*. Throughout the paper, whenever the notion of goal is used, it will be meant as an achievement goal in the above sense, unless otherwise specified.

Indeed, in our model (as well as in Cohen and Levesque's), an achievement goal is not yet an intention.

Cohen and Levesque's theory includes a notion of relativised goal:

$$\begin{aligned}(\mathit{R\text{–}GOAL}\ x\ p\ q) \overset{def}{=}\ & (\mathit{A\text{–}GOAL}\ x\ p) \wedge \\ & (\mathit{BEFORE}((\mathit{BEL}\ x\ \neg q) \vee (\mathit{BEL}\ x\ p) \vee (\mathit{BEL}\ x \neg \Diamond p)) \\ & \quad \neg(\mathit{A\text{–}GOAL}\ x\ p)) \qquad (5)\end{aligned}$$

x has a goal p relativised to q, when x has an achievement goal p, and before ceasing to have p as an achievement goal, x believes either that p is realised or unachievable or that the escape condition q does not hold. Essentially, this means that x has p as long as and because he believes that q.

Our notion of a goal (Conte and Castelfranchi 1995a) is slightly weaker than that allowed by Cohen and Levesque. We propose to treat goals as *realistic* desires, rather than *chosen* ones. In our terms, a goal is but a regulatory mental attitude which calls for a series of operations, including some preliminaries, involved in planned action. In other words, along the lines of classical AI planning systems, we define a goal as a device which activates planning and *action*. In our terms, a goal may be abandoned not only when it is believed to be fulfilled or unachievable, but also when it is found incompatible with another more important goal.

The pred *OUGHT* intuitively means that there is some sort of *obligation* on proposition p. For the time being, we take it as an atomic one-place predicate, although it seems possible to further analyse it as some sort of external reason which forces a given goal, namely the adoption of a given goal. However, we will assume obligation as a primitive, which defines a set of worlds in which p follows from obligations. The relation of accessibility ***O*** is a subset of ***B***.

In our model, agents have normative beliefs when they think there is an obligation on a given set of agents to do some action.

In the following, x and y denote agent variables with $x \neq y$ always implicitly stated, a denotes an action variable, e a sequence of events, r a resource, and p and q well formed formulae representing states of the world (with $p \neq q$).

We express the general form of a normative belief as follows:

$$(\mathit{N\text{–}BEL}\ x\ y_i\ a) \overset{def}{=} (\Lambda_{i=1,n}(\mathit{BEL}\ x(\mathit{OUGHT}(\mathit{DOES}\ y_i\ a)))) \qquad (6)$$

in words, *x has a normative belief about action a relative to a set of agents y_i if and only if x believes that it is obligatory for y_i to do action a*. The predicate *OUGHT* here stands for an *obligation for a set of agents y_i to do action a*. A few words are needed to elucidate the semantics of our predicate *OUGHT*. This stands for an operator of obligation about any given state of the world. However, it should be taken in a somewhat weaker sense than what is usually meant by obligation in traditional deontic logic. In fact, while in traditional deontic systems, p necessarily follows from obligation (that is to say, it is not possible that at the same time p is false and obligatory), in other systems (Jones and Pörn 1991), two concepts need to be distinguished, one referring to deontic necessity and the other to another type of obligation. The latter is defined as the circumstance in which a given proposition is both obligatory and possibly false in some sub-ideal world.

In order to express normative goals, another belief is needed, namely a pertinence belief: for x to believe that he is addressed by a given norm, he needs to believe that he is a member of the class of agents addressed by that norm:

$$(P\text{–}N\text{–}BEL\ x\ a) \overset{def}{=} (\Lambda_{i=1,n}(N\text{–}BEL\ x\ y_i\ a)) \wedge (V_{k=1,n}(BEL\ x(x = y_k))) \quad (7)$$

where *P–N–BEL* stands for normative belief of pertinence; in words, *x has a normative belief of pertinence when he has a normative belief relative to a set y_i and an action a, and believes that he is included in y_i.*

Now, x's beliefs tell him not only that there is an obligation to do action a, but also that the obligation concerns precisely himself.

We have not seen any normative goal yet. A normative goal is defined here as a goal always associated with and generated by a normative belief. Let us express a normative goal as follows:

$$(N\text{–}GOAL\ x\ a) \overset{def}{=} (R\text{–}GOAL\ x(DOES\ x\ a)(P\text{–}N\text{–}BEL\ x\ a)) \quad (8)$$

or, *x has a normative goal concerning action a when he has the goal to do a relativised to his pertinence normative belief concerning a*. A normative goal of a given agent x about action a is therefore a goal that x happens to have as long as he has a pertinent normative belief about a. Ultimately, x has a normative goal in so far as he believes that his is subject to a norm.

2.2. HOW MANY GOALS BEHIND ONE NORM?

Let us analyse the imperativistic view (see above §1.2). There are several goals concealed, so to speak, under one and the same norm. Let us try to disentangle them.

The first goal implied by a norm is the goal that a given state of the world is produced thanks to a norm conformity. If a is the action prescribed by a norm, the imperativistic view is bound to assume that there is an agent x wanting the state of the world[5] consequent to the compliance with a

$$\forall a(((DONE\text{a}) \supset p) \wedge (OUGHT(DONE a))) \supset \exists x(GOAL\ x\ p) \quad (9)$$

[5] Of course, such a state may be a more or less arbitrary consequent of the norm conformity. Often, the function of conventions is relatively independent of the specific content of the convention in question (as in the case of the norms of precedence).

Let us call this goal of x's the *raison d'etre* of a, its expected utility. But a further goal may be implied by the prescriptive view, namely,

$$\forall a(OUGHT(DONEa)) \supset \exists x(GOALx(DONEa)) \qquad (10)$$

in words, according to such a view of norms, if there is a norm, there must be someone wanting it to be fulfilled.

Indeed, other goals may be supposed to be implied by a norm,

$$(\Lambda_{i=1,n}(OUGHT(DOES\ y_i\ a)) \supset \exists x((GOALx(DOES\ y_i\ a)) \qquad (11)$$

which reads, if there is a conjunction y_i of agents from 1 to n which has the obligation to do a, there must be an agent x who wants y_i to do a. Furthermore,

$$(OUGHT(DOES\ y_i\ a)) \supset \exists x(GOALx(N\text{–}GOAL\ y_i\ a)) \qquad (12)$$

that is, an obligation whatsoever implies someone wanting the obligation to be fulfilled by those people who ought to fulfil it. But in order to have this goal, y must know what x wants of her:

$$(OUGHT(DOES\ y_i\ a) \supset \exists x(GOALx(BEL\ y_i(GOALx(N\text{–}GOAL\ y_i\ a))))(13)$$

Finally, in an imperativistic view, a norm necessarily implies that there is someone wanting it to be prescribed:

$$(OUGHT(DONEa) \supset \exists x\exists z(GOALx(OUGHT(GOALz(DONEa)))) \qquad (14)$$

where z does not necessarily coincide with x. The last goal, indeed, is that which most explicitly defines the issuing of norms. In the imperativistic view, a norm is there if there has been the goal of issuing it.

Now, as said above, such a view is exceedingly strong. We are likely to propose a view of norms that is grounded on a general notion of goal, but does not imply (14). Getting rid of (14) allows us to maintain a prescriptive view of norms that accounts for conventions *on condition that*:

(i) the emergence of conventions be not viewed as a continuous process, such that the higher the conformity to regularities, the more valid the resulting norm; (ii) an essential step in the above process, however implicit, is found to be a cognitive one. But in our terms, unlike the imperativistic stance, such a cognitive step is not seen from the point of view of the legislator, that is to say, the agent issuing the norms. Rather, we will take the complementary view, namely that of the addresse. We will claim that *a norm is there whenever there is someone believing that there is someone who is prescribed to comply with that norm*. We think this allows Hart's intuition to be made explicit. By recognizing a given behaviour as not only generally expected, but also as wanted, people give grounds, pave the way, and ultimately "issue" the norm. In §3, we will see this view more closely.

2.3. OBLIGATIONS AND MENTAL STATES. A POINT MISSED IN THE FORMAL TREATMENT OF NORMS

Indeed, volitions, or more generally, wants are necessarily implied by norms, according to the imperativistic view. But, on the other hand they are insufficient. In Kelsen, obligations imply norms issued by an act of normative production. But the latter implies a meta-norm, and not simply anyone's want. This imples that only a subset of the agents are held to issue norms. And only a subset of these agents' wants can enforce norms. What is the link between between such wants and enforced norms?

This is a rather general problem. Generally speaking, there is not much formal theory of the connections between goals and obligations (see Shoham and Cousins, 1994). The formal models of mental states are not able to acoount for these connections.

First, both in the BDI architecture (that is, the model of agency in terms of Beliefs, Desires and Intentions – cf. Rao et al., 1992 – and currently used for implementing Multi-Agent Systems) and in Cohen and Levesque's model (1990), goals may only arise from desires. Now, by definition, an obligation is a non-desirable state: if a worldstate were desirable, there would be no need for an agent to be obliged to obtain it (see the notion of exogenous motivation in Elster, as quoted by Bicchieri, 1990). Hencefore, within the current architetcures of mental states, the relation between obligations and goals is impossible. Their intersection is empty. Consequently, obligations are primitive. In his turn, Shoham (Shoham & Tennenholz, 1992) considers them as constraints reducing the agent's set of available actions, rather than as goals.

Secondly, in the BDI architecture, goals are vanished. They have been replaced by intentions. However, the notion of intention is much more restrictive than that of goal. This constraint plays a fundamental role with regard to obligations. In fact, intentions are decided upon goals. They are a subset of those desires which are belief-compatible. But an operational question comes out here: since the starting operator is an operator for beliefs, which by default is not regulatory, how is the whole system triggered? How are belief-compatible desires selected for? A goal, in its more abstract role, is simply an internal state which puts the system on-line, so to speak, making it check whether that state is to be achieved or not. But if one such state is not active, how is the system activated? Analogously, how is it possible for a system to accept, to decide whether to comply with one obligation if there is no corresponding goal? How can obligations give rise to intentions, if an intention is an already decided upon desire?

In our terms, instead, a goal is a rather general notion. A goal is but an internally represented state that acts as a regulatory state. Indeed, a goal is *the* most general mechanism of cognitive regulation.[6] However, in our terms, a goal is not a decided upon goal. On the other hand, a goal is not necessarily a subset of desires. An obligation, as well as a request by some other agent, may give rise to a goal thanks to means/end reasoning. Obviously, this does not mean that a goal will be necessarily formed starting from an assumed obligation. A fortiori, a goal generated by an obligation will not necessarily lead to an intention. It may be abandoned for a number of reasons (it is already fulfilled, it is impossible to achieve, it is incompatible with more important goals). But, in order to check whether to achieve a wanted state, the system must be activated, so to speak, by that state. This is equal to saying, that in order for an intention to be formed, a system must be regulated by a goal. In our model of norms as mental objects (cf. Conte & Castelfranchi, 1995b), we attempt to find some bridging mechanims between obligations and goals. We have defined normative goals as goals motivated by normative beliefs.

One could say that for the scientists of law it is irrelevant to explore the link between the legal domain and others. This paper proposes a radically different view. Such link seem to be fundamental because (a) it is one of the issues debated by many philosophers of law (think of the socalled "realistic" school of thought); (b) the link between spontaneous and deliberate norms may help understand the role of spontaneous social control in the spreading of norms, whether legal or social. People control one another's compliance with both laws and conventions. Spreading mechanisms, and especially the social control, operate much in the same way (I will put the blame on you both if you ignore the speed limits and if you eat with your hands). (c) The rules and mechanisms for reasoning about norms and to some extent also for deciding whether to comply with them are essentially independent of whether what is reasoned and decided upon is a legal or a social norm. Specific reasons for transgression and obedience may differ but the deciding algorithm and the representational format are essentially the same.

3. Believed Prescriptions

A unifying theory of norms must be grounded upon a notion of prescription that does not imply deliberate issuing. Such notion of prescription is conceivable only if one takes the complementary perspective to that implied by the imperativistic view, namely the viewpoint of the would-be conformer. In other terms, a social norm here is tentatively defined as an *implicit prescription* nested in the conformer's normative beliefs. To state it differently, a social norm is there whenever there is someone believing that a given behaviour is prescribed of a given population (subset):

$$\forall a (N\text{–}BEL\ x\ y_i\ a)) \supset (BEL\ x \exists z (GOAL\ z (DOES\ y_i\ a))) \qquad (15)$$

[6] In this perspective, we do not agree with those analyses of second-order operators, such as action, purpose and will (cf. Holmström-Hintikka 1991) as irreducible to one another. In our view, an action is reducible to an intention, which in turn is reducible to a goal.

behaviours may be perceived as prescribed without having being explicitly issued. Obviously, a norm is there even when its addressees do not recognize it, provided there is someone who has the corresponding mental representation. A complicated legislature which is not understood by the people is a norm. A norm is not necessarily good, nor shared by its addressees. If people refuse to pay the taxes, the norm prescribing that taxes must be payed is still in force but is probably poorly efficaceous, and possibly contrasting with other norms (may be social, or moral) the agents are (more or less instrumentally) referring to.

But some other specifications are required

(i) who is z perceived to be? The imperativistic view would answer that z is a normative authority. But this is a tautology, as was shown above. Another answer is that of saying that z is a subset of the same community which y_i belongs to. Therefore, $z \cap y \neq \emptyset$. The larger the intersection in x's assumptions, the more x believes the norm is in force, and the more mandatory the norm is.

(ii) (15) is still insufficient since z is perceived to simply *want* that y_i does a. No duty has been founded yet. Therefore (15) should be replaced by

$$\forall a(\textit{N–BEL}\ x\ y_i\ a)) \supset (\textit{BEL}\ x \exists z(\textit{GOAL}\ z(\textit{N–GOAL}\ y_i\ a))) \qquad (16)$$

which expresses the idea that any obligation is such that there is someone believing that a given subset of the population z wants the obligation to be fulfilled by a set of agents y within the same population as long as y believes to undergo that obligation. But what are the grounds of such obligation? How is it formed within the agents' minds? In a forthcoming paper, we will provide some preliminary answers to this question.

In this paper, we will examine the effects of (16) on the spreading of an obligatory action.

4. Issuing Social Norms while Adopting Them

The decision to conform to what is perceived to be an obligation plays a relevant role in its spreading over a population of cognitive agents. While the conventionalistic view derives social norms from the spreading of conformity, here conformity is derived, so to speak, from the spreading of obligation-recognition and -adoption. The very act of accepting an obligation implies and turns into enforcing it. The agent respecting the obligation turns into a supporter. Conforming leads to prescribing. The agent undergoing an obligation becomes a legislator. The more an obligatory behaviour is believed to be prescribed, the more it will be complied, and the more, in turn, its prescription will be enforced. Rather than acting only through a behavioural contagion, a passive social impact, the spreading of norms is affected by cognition in a variety of ways:

(i) *it leads to implementing effective conformity*. If the number of conformers was c_i before x had realised an obligatory action is prescribed, now that x realised it, chances are that the number will be increased by 1 unit. This is equivalent to the effect of imitation.

(ii) *Effective conformity contributes to the spreading of normative beliefs*. The larger the number of conforming agents and the more likely the observers will form normative beliefs. However, the formation of normative assumptions is conditioned to a number of contextual clues which activate the inferential mechanisms mentioned above. In other terms, the spreading of normative conformity is not a self-enforcing mechanism, as the conventionalistic approach seems to assume. It is necessarily mediated by cognitive mechanisms.

(iii) *The spread of normative beliefs contributes to the spreading of normative actions.* The wider the spread of normative beliefs, and the higher the chances that conformity is due to the formation of normative goals (cognitively, a goal relativized to the belief that one is addressed by a given norm; formally, a goal arising at the intersection between goals and normative beliefs; for a treatment, see Conte & Castelfranchi, 1995b). This is also allowed by some reasoning mechanisms and rules, for example reciprocation (if I believe to benefit from a given behaviour, I will be more likely to display the same behaviour as long as I believe it is generally prescribed that benefits should be somehow returned).

(iv) *The spread of normative actions contributes to the spreading of normative influence.* The larger the number of agents conforming to one given n, and the more distributed will be the want that other agents will conform to the same norm. This is due to:

- an *equity* rule (cf. Conte & Castelfranchi, 1995a, ch. 7). People do not want others in the same conditions as their own to sustain lower costs – benefits being equal (this is, indeed, one the most probable explanations of the Heckathorn's (1990) group sanction control: the more agents respect the norms, and the more likely they will be to urge others to do the same).
- "*norm-sharing*". Agents are likely to "share" the respected norms, that is, to believe that those norms are sensible, useful, necessary, etc.. This is in part a real phenomenon (agents decide to comply with the norms they share). But it is also a powerful self-defensive mechanism (agents share the norms they happened to respect). In both cases, agents will defend the norms they share (again, for a formal notion of norm-defending goal, cf. Conte & Castelfranchi, 1995b), implementing the number of agents who want those norms to be respected.

(v) *The spread of normative influence contributes to the spreading of normative beliefs*, and the whole process is started again in a circular way. This type of model, evidently lends itself to validation through computer simulation.

It is interesting to observe that the same line of reasoning applies to the reverse situation: it might be interesting to explain how the mechanism described above norms may become deadletter, that is, how transgression, as well as obedience, propagates thanks to the spreading of the negative influence of the other agents subject to the same norm.[7]

5. Conclusions

In this paper, we have addressed the issue of a unified view of norms. A short review of some formal, philosophical and social scientific literaure on norms suggested that norms have received a dychotomic treatment. In fact, while the rational view has privileged the study of the emergence of conventions form mere regularities, thus undermining the prescriptive view of norms, philosophers of law have emphasized the deliberate issuing of norms, thus disregarding conventions and customary, implicit, social norms.

Consequently, an attempt is made to fill the gap between these views. The process from conventions to prescriptions is defined as a non-continuous process, in which cognition plays a fundamental mediatory role. A bridge is identified in a general notion of distributed goals. Rather than focusing on the issuing of prescriptions, the complementary perspective has here been taken. A social norm is seen to imply a belief that a given behaviour is generally prescribed within the population observed. A given behaviour is executed because and as long as it is believed to be obliged. As a consequence, the act of conforming to a given conduct – as long as it is believed to be prescribed – gives rise to the act of prescribing it, thus contributing to its spreading. Some cognitive mechanisms responsible for the spreading of norms are finally examined.

Acknowledgements

We would like to thank Prof. Giovanni Sartor for his carefully reading of this paper and for his invaluable suggestions, as well as the anonymous reviewers of the 2nd Workshop of the ModelAge group, who usefully commented on a previous draft of this work.

References

Axelrod, R. 1984. *The Evolution of Cooperation*, New York: Basic.

Bicchieri, C. 1990. Norms of cooperation. *Ethics* 100, 838–861.

Cohen, P. R. and Levesque, H. J. 1990. Persistence, Intention, and Commitment. In P. R Cohen, J. Morgan and M.A. Pollack (eds.), *Intentions in Communication*, Cambridge, MA: The MIT Press, 33–71.

Conte, R., Miceli, M., and Castelfranchi, C. 1991. *Limits and Levels of Cooperation. Disentangling Various Types of Prosocial Interaction.* In Y. Demazeau, and J. P. Mueller (eds.), *Decentralized AI-2*, Armsterdam: Elsevier, pp. 147–157.

[7] We would like to thank one of the anonymous reviewers who have commented upon a preceding version of this paper for such a useful remark.

Castelfranchi, C., Miceli, M., and Cesta, A. 1992. *Dependence Relations among Autonomous Agents.* In Y. Demazeau, E. Werner (eds.), *Decentralized AI-3*, Amsterdam: Elsevier, pp. 215–231.

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

Conte, R. and Castelfranchi, C. 1995b. From normative beliefs to normative goals. In J.P. Mueller & C. Castelfranchi (eds.), *From Reaction to Cognition*, Berlin: Springer, pp. 186–199.

Conte, R. 1994. Norme come prescrizioni: per un modello dell'agente autonomo normativo. *Sistemi Intelligenti* 1, 9–34.

Halpern, J. Y. and Moses, Y. O. 1985. *A Guide to the Modal Logics of Knowledge and belief. Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Los Altos, CA: Kaufmann, pp. 480–491.

Hart, H. A. 1961. *The Concept of Law.* London: Oxford University Press.

Heckathorn, D. D. 1990. Collective sanctions and compliance norms: A formal theory of group-mediated social control. *American Sociological Review* **55**, 366–383.

Heiner, R. A. 1983. The origin of predictable behavior. *American Economic Review* **73**, 560–595.

Heiner, R. A. 1986. Imperfect decisions and the law: On the evolution of legal precedent and rules. *The Journal of Legal Studies* **15**, 227–262.

Holmstrom-Hintikka, G. 1991. Action, purpose and will. *Acta Philosophica Fennica* special issue, 50.

Jennings, N. R. 1993. Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review* **3**, 223–250.

Jones, A. J. I. and Porn, I. 1991. *On the Logic of Deontic Conditionals*. In J. J. C. Meyer and R. J. Wieringa (eds.), *First International Workshop on Deontic Logic in Computer Science*, pp. 232–247.

Kelsen, H. 1934/1979. *Allgemeine Theorie der Normen*. Wien: Monasche Verlag-und Universitaets-Buchhandlung; Eng. tr. *A General Theory of Norms*. Oxford: Clarendon Press, 1991.

Kerr, N. L. 1995. Norms in social dilemmas. In D. A. Schroeder (ed.), *Social Dilemmas. Perspectives on Individuals and Groups*, London: Praeger, pp. 31–49.

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

Rao, A. S., Georgeff, M. P., and Sonenberg, E. A. 1992. Social plans: A preliminary report. In E. Werner and Y. Demazeau (eds.), *Decentralized AI-3*,. North Holland: Elsevier, pp. 55–77.

Ross, A. 1958. *On Law and Justice*, London: Steven & Sons.

Schelling, T. C. 1960. *The Strategy of Conflict*, Oxford: Oxford University Press.

Schotter, A. 1981. *The Economic Theory of Social Institutions*, New York: Cambridge University Press.

Shoham, Y. and Cousins, S. B. 1994. Logics of mental attitudes in AI. In G. Lakemeyer, B. Nabel (eds.), *Foundations of Knowledge Representation and Reasoning*, Berlin: Springer.

Shoham, Y. and Tennenholtz, M. 1992. On the synthesis of useful social laws for artificial agent societies. *Proc. of the AAAI Conference*, Stanford, CA: The American Association of Artificial Intelligence, pp. 276–281.

Ullman-Margalit, E. 1977. *The Emergence of Norms.* Oxford, Oxford University Press.