Artificial Intelligence and Law **7:** 1–15, 1999.

Introduction: Agents and Norms: How to fill the gap?

ROSARIA CONTE and RINO FALCONE
Division of AI, Cognitive and Interaction Modelling Institute of Psychology, CNR, Rome, Italy
E-mail: rosaria,falcone@pscs2.irmkant.rm.cnr.it

GIOVANNI SARTOR
School of Law, Queen's University of Belfast, Belfast BT7 1NN
E-mail: gsartor@qub.ac.uk

1. Two Frameworks for Norms and Agency

AI has so far approached normative concepts and phenomena especially in the two following frameworks:

- Theory of the law and related computational applications, especially in the areas of legal expert systems, normative reasoning and diagnosis, etc.;
- Theory of multi-agents systems (MAS) and related computational applications, especially in the areas of Computer Supported Cooperative Work (CSCW), electronic commerce, etc.

A wide gap exists between these two frameworks. They differ in terms of

- *language and formalisms used* (mostly logic-based in the legal domain and more oriented to implementation languages in the multi-agent domain);
- *theories of reference* (philosophy of law and deontic philosophy in the former domain, as opposed to agent theory and game theory in the latter);
- *objectives* (models of legal institutions, legal information systems, in the former, as opposed to social theory and optimization of coordination and cooperation in the latter);
- *underlying philosophy* and concept of a norm (mainly interpreted in the legal, institutional sense in the former, and as a social, customary norm or convention in the latter).

We believe that an approach to norms and agency capable of meeting the requirements of the emerging field of autonomous agents requires integrating the

results obtained in the legal and in the multi-agents domains. In this introduction, we will

- summarize the different approaches to norms adopted in the two domains (Section 2);
- formulate some open questions and argue that their solution requires a synthesis of those approaches (Section 3);
- present the papers included in this volume as attempts to answer some of these open questions (Section 4).

2. Agents and Norms in Legal Theory and in Multi-Agents Systems

Both legal theory and multi-agents theory have worked out models of norms and agency which are appropriate in regard to some specific objectives. However, we shall argue that neither of those discipline has been capable of providing a link between norms and agency which is appropriate for autonomous agents, i.e. agents than can adopt normative attitudes (beside other attitudes) and can view those attitudes as (defeasible) reasons for their behaviour.

2.1. THE POINT OF VIEW OF LEGAL THEORY

Most problems concerning regulation of the interaction of autonomous agents are linked to issues traditionally addressed by legal studies, and specifically, by legal doctrine and legal theory (we do not consider here more marginal legal disciplines, such as legal anthropology or legal sociology). This is no surprise, since law is the most pervasive and developed normative system, and it is typically concerned with the government of autonomy: the fundamental task of the law is exactly that of providing normative reasons which may restrain and co-ordinate the behaviour of autonomous agents, each one of whom "can use his own knowledge for his own purposes" (Hayek 1973, 55).

It would be impossible to list here all problems where the student of autonomous agents can find useful suggestion in the law (the definition of normatively protected domains of action, the establishment of mechanisms for delegation and representation, the forms and procedures for engaging into binding agreements, the conditions and the consequences of the ascription of responsibilities, etc.). It is true that the law rarely provides models which are so clear-cut and formally refined that they can be directly transferred into a computable representation, but the importance of the fact that the law can provide workable solutions to coordination problems, tested both in the doctrinal dialectics and in the legal practice, could hardly be underestimated.

However, to profit from the contributions of legal disciplines, we must also acknowledge the limitations of current legal approaches: the law offers various

ingredients for a theory of the norm-governed interaction of autonomous agents, but does not provide an adequate integration of those ingredients. In particular, an adequate model of such an interaction requires the combination of mental, social and normative components, which in legal studies have usually been investigated in complete separation, under different disciplines, and in regard to different questions.

Mental notions are usually considered in the framework of specific legal theories, where those notions are linked to the factual determinants of agency rather that to its normative components. For example the doctrine of criminal liability provides a learned and detailed discussion of the concepts of the intention and will of performing a certain action, and of how intention and will are influenced by the knowledge of the effects of that action (e.g. are the effects which are foreseen, but not willed, by the actor to be considered as intended by him/her?). Similarly, in the doctrine of contracts, we can find endless discussion concerning the role of the intention of the parties in determining the effects of the contracts (are all consequences of a contract to be intended by both parties, does the declaration prevail over their intention, what is the relevance of (erroneous) beliefs of the parties?). In both doctrines, on the contrary, normative beliefs have received a limited attention, since the effects established by the relevant legal norms will usually follow regardless of the knowledge of those norms by the addressees.

Within legal studies, the basic normative notions are usually considered in the framework of legal theory, which provides numberless accounts of the nature of legal norms, of their sources, of their typology, of their combination in the legal system, of the positions of their addresses, the reasoning processes leading to the application or to the acceptance of norms. However, also in legal theory the mental dimension of norms has been given a very limited attention. This is especially true for those (still dominant) approaches with are inspired by the positivistic identification of the law with what is prescribed by the political authorities (or by the State). From this perspective, the focus is on the author of the norm, rather than on its addressees, which induces a “voluntaristic” approach to the law: the basic (or the only) mental notion to be considered is that of the will of the political authorities or of the organisation which unifies them, the State.

For example, the greatest representative of legal positivism, Hans Kelsen ends up by affirming that legal norms are meanings of acts of will (no imperative without an emperor) (Kelsen 1979). From such a point of view, a consideration of the psychological attitudes of the norm addressees would be out of place, and they would indeed express that commixture of legal and psychosocial elements which is excluded by the ideal of a “pure” science of law. The dynamic of a normative system is indeed reduced to the issuing of new prescription by normative authorities (according to higher level norms) while the spontaneous spreading of normative attitudes plays no significant role. Correspondingly, the role of customary rules, whose existence is based on the attitudes of the addressee (traditionally, two ele-

ments are distinguished in a custom, a constant behaviour, plus the opinion of the obligatoriness of that behaviour) is downplayed.

A very restricted view of the psychological dimension of legal norms can be found in the English version of legal positivism (the so called analytical jurisprudence). In particular, Jeremy Bentham and John Austin understood law as being the command of the holder of political power, supported by the threat to visit with sanction those who violate that command. From this perspective, the only mental counterpart of the will of the political authority is the fear of the norm addressee, which compels him/her to respect the norm. Spreading of the norms is just a side-effect of the expansion of the political power of the authorities issuing them (and of their possession of the means through which they can visit their subjects with the threatened sanctions).

Also the other major tradition of legal thinking (and the main antagonist to legal positivism), that is natural law school, has paid little attention to the mental dimensions of legal norms. This is basically due to the fact that, at least according to the predominant rationalistic strand of the natural law school, the focus is on rational acceptability (intended, as the derivability from rationally evident axioms) of norms. The mental attitudes of the addressees of legal norms and the reasoning processes which determine those attitudes can be left aside since the acceptance of a norm as binding can be considered as a straightforward consequence of its evident rationality (although the doctrines of social contracts can possibly be understood as an idealisation of the social process leading to the acceptance of legal norms).

To find some analyses of the formation and spreading of norms which pay due attentions to the normative attitudes of norm addressees (so as to allow us to establish a link with AI contributions on autonomous agents) we need to consider some other trends in legal thinking.

The point of view of the addressees of legal norms is explicitly adopted, for example, by the so called Scandinavian realists (Alex Hägerström, Karl Olivecrona, Alf Ross, etc.). Those authors viewed the foundation of the law in the normative beliefs of the addresses of legal norms, and provided an analysis of some aspects of the psychology of norm addressees, and especially of the sense of duty which is linked to the acceptance of a norm. From their perspective it is the psychological attitude of the citizens which establishes the validity of legal norms, even of the highest ones: the opinion of the citizens (their conviction that the legislator is entitled to issue norms, in certain domains and for certain purposes) grounds and limits the very power of the legislator. However, those authors viewed normative beliefs to be based on a necessary mistake: the attitude of the norm addressees (their acceptance of certain norms) is only explicable as the result of an erroneous belief in the extra-empirical existence of an imaginary kingdom of norms. Correspondingly they tended to exclude any relevance to rationality and reasoning in the processes of norm formation and spreading, and mostly limited their contribution to a critical analysis of “normative ideologies” (those mistaken beliefs in the existence

of norms), rather then going into a logical analysis of normative attitudes and of the mental processes leading to their adoption.

A different insight into the processes of norm formation and norm spreading can possibly be obtained from those approaches which focused on the evolution of customary norms, such as the historical school in Germany, which saw norm formation as the manifestation of the development of the "spirit of the people", and especially the Scottish social philosophy of David Hume and Adam Smith, which advanced a "conventional" view of norms formation (norms emerge for their capacity to provide effective solutions to problems of social co-ordination). The intuitions of these last authors has been recently developed into an evolutionary theory of the formation and spreading of normative attitudes by Friedrick Hayek (1973), according to whom social norms are selected, by evolutionary processes, for their capacity to sustain a successful spontaneous order, which integrates the actions of individuals and makes their autonomous use of distributed knowledge possible. This same tradition is developed in a different way by Gerard Postema (1982), who, instead, emphasises rational choice and models norm formation as the solution of a co-ordination problems, to be understood according to game theory.

Finally, the role or reason and rationality in the acceptance and the spreading of norms is illustrated by the various theories of legal reasoning (McCormick 1978; Peczenick 1989) and on the role of norms as reasons for actions (Raz 1975). Those contributions which focus on the notion of authority (Raz 1986, 23ff) help us in understanding the process through which a norm issued by an authority may be accepted as (legitimately) binding by his addressees, i.e., how the power of the authority (and its products) may be deemed to be justified. Moreover, those approaches which focus on the dialectical features of legal reasoning, help us to understand how norms may be selected via dialectical procedures, and how they may impact on those very precedures (Alexy 1989).

From the sketchy discussion just provided, we may observe that, although the whole mental dimension of normative attitudes and norm formation has not been adequately investigated in the legal disciplines, many insights and suggestions can be derived from legal studies. Still, those insights and suggestions have very rarely being cast in a formal framework, as required for the development of computable representation. In this regard, legal theory may exhibit the various results obtained in the formalisation of normative concepts, from deontic-logic (cf. for all Alchourrón & Bulygin 1971) to the theory of normative position (Lindhal 1977), to the dynamics of normative systems (e.g., the theory of belief revision of Alchourrón, Gärdernfors and Makinson (1985) was originally intended to address legal concerns). In particular, those studies have provided us with a formal analyses of norms which is based on the distinction between ideality and reality (Jones and Pörn 1985), and emphasises the possibility that norms are violated. However, legal theory has not yet made the step from the formalisation of legal conceptions (cf., Herrestad 1995) to a full fledged representation of the mental attitudes towards those conceptions and of their dynamics.

Interestingly, although modal logic has been extensively used to model deontic and normative notions, the idea to use it as a comprehensive framework for normative attitudes (Pörn 1977) has not yet been received into legal theory. The same Carlos Alchourrón, who combined better than anybody else a mastership of logic and legal theory, adopted mainly a positivistic attitude, considered norm formation as the result of the activity of the legislator, and centered his formal model of norm formation on the latter (cf., Alchourrón 1991).

The formalisms provided by legal theory have been developed and enriched in the framework of the AI & law, where they have been combined with the resources of AI & computer studies. It would not be possible to provide here an account of the multifarious AI & law proposal in the representation of norms and normative reasoning. For our purposes, it is sufficient to remark that usually mental notions (what minds know/accept what norms and facts, how this acceptance is brought about, how it is linked to other psychological attitudes) have be usually left implicit also in AI & law research. The accent has indeed been in the use of a unique store of knowledge in legal problem solving (as in traditional expert systems), rather than in the dynamic interaction of multiple agents, expressing different points of view.

Only recently, especially in the development of formal models of legal argumentation the different attitudes of the agents involved in a legal debate have been to a certain extent formalised (cf., for all, Gordon 1995) and implemented (Nitta 1995). However, the problem of an explicit formal representation of agent's normative attitudes (rather than of normative contents) and of the development of formal models of their formation been tackled to a very limited extent also within AI & law research.

2.2. THE POINT OF VIEW OF MULTI-AGENT THEORY

The advent of large communication networks, civic networks, as well as the spread of electronic commerce, contributed dramatically to draw the attention of the AI scientific community to various normative 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). More specifically, the efforts done by MAS researchers and designers to construct *autonomous* agents (Wooldrige & Jennings 1995) carry with themselves a number of interesting but difficult normative issues:

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

(c) More generally, how to obtain a robust performance in teamwork (Cohen & Levesque 1990)? How to prevent agents from dropping their commitments, or better, how to prevent agents from disrupting the common activity (cf., Jennings 1992; Kinny & Georgeff 1994; Singh 1995)?

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

The Multi-Agent study of norms draws upon the treatment of norms in the social sciences, and tends to adopt a rationalistic approach: norms are often viewed as *emergent* properties of utilitarian agents' behaviour. Therefore, in the Multi-Agent field, social norms are perceived as devises to help improve coordination and cooperation (Shoham & Tenneholz 1992; Jennings and Mandami 1992; Conte & Castelfranchi 1995; Walker & Wooldridge 1995). However, in both areas no adequate representation of the agent's internal states, including normative attitudes, is adopted (this parallels the insufficient attention for the psychological aspect of norms that we remarked in regard to legal theory). This affects the way in which three questions, of vital importance are addressed:

1. How do agents acquire norms?
2. How can agents violate norms?
3. How can the agent be autonomous?

Let us first consider normative innovation. In the field of formal social science,[1] the spread of norms and other cooperative behaviours is usually not explained by modelling internal representations of norms. The object of inquiry usually consists of the conditions under which agents converge on behaviours which prove efficient in solving problems of coordination (Lewis 1969) or cooperation (Axelrod 1987), independently of the agents' beliefs and goals (Binmore 1994): no theory of the acquisition of normative attitudes as grounded upon agents' internal representations has yet been provided. This is also true in multi-agent systems, where norms are exsplicitly represented in the agents, but only as built-in constraints. This

[1] That is, in utility theory and in game theory. Social (psychological) theorists have attempted behavioural explanations of normative influence. 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 psychologists pay attention to rules of reasoning (natural vs. formal logics) rather than to moral and social norms. Generally speaking, the role of cognition in social action is still relatively poorly explored.

means that the connections between obligations and mental states are theoretically overlooked, and usually not formalised (Shoham & Cousins 1994).

This strongly limits the results that can be achieved in the very important area of normative learning in multi-agent systems, which has not only a theoretical, but also a practical relevance. If agents are enabled to acquire new norms, there is no need for expanding exceedingly the knowledge-base of individual agents. Consequently, the multi-agent system may be optimized when it is *on-line*, while multiagent systems where norms have been hardwired into the agents allow for a modification of norms only when the system is *off-line* (Shoham & Tennenholz 1992). In a successive work, indeed, these authors have introduced the notion of co-learning,[2] which refers to a process in which several agents simultaneously try to adapt to one another's behavior so as to produce desirable global system properties. Of particular interest are two specific co-learning settings, which relate to the emergence of conventions and the evolution of cooperation in societies, respectively. Shoham and Tenneholtz have defined a basic co-learning rule, called Highest Cumulative Reward (HCR), which gives rise to nontrivial system dynamics. The study shows the eventual convergence of the co-learning system to desirable states, and the efficiency with which this convergence is attained. Results on eventual convergence are analytic: the results on efficiency properties include analytic lower bounds as well as empirical upper bounds derived from rigorous computer simulations. The same result has been achieved by Walker and Wooldridge (1995) in their simulation study about the emergence of conventions in multi-agent systems.

Despite the indubitable significance of the results just mentioned, we think that the treatment of norms as action constraints cannot answer some important questions as to how norms emerge. In particular, the treatment of norms as emerging conventions resulting from co-learning processes, can only deal with how pre-existing actions are gradually generalised or dropped. It cannot explain the process of the acceptance of new norms stated by an authority (a phenomenon which, as we have seen, has been emphasised in legal theory), and more generally those cases where a norm is selected which prescribes an action which nobody practiced before. Norms not only constrain an agent's conduct, making it more uniform and predictable, but they do also provide new behaviours (e.g., pay the taxes; wear a helmet while driving a motorcycle, etc.).

Besides its incapacity of dealing adequately with norm acquisition, the current treatment of norms in multi-agents systems shows its limitation also in its inability of dealing with violations. As shown above, in the MA field, norms are treated as constraints to either the agent's action repertoire (Shoham & Tenneholz 1992; Jennings 1995) or its evaluation module (see Boman's paper in this issue). Norms operate by reducing the set of available or convenient actions 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

[2] Shoham, Y. and Tennenholtz, M. *Co-Learning and the Evolution of Social Activity*, CS-TR-94-1511.

conflicts of norms, which often arise among tasks associated with different roles, or among norms belonging to different domains of activity. This feature is crucial with regard to both legal expert systems and autonomous agents interacting in a common world.

The incapacity of acquiring new norms and of violating norms seriously limits the agents' autonomy, which limitation has considerable practical implications. If we need fully autonomous agents, we also need autonomous normative agents: a capacity for autonomous norm-acceptance would greatly enhance multi-agent systems' flexibility and dynamic potentials. An autonomous normative agent has an increased selective capacity (potential for selecting those external requests which it is necessary or convenient for it to fulfil), which encompasses not only norm-applying *decisions*, but also the acquisition of new norms. Indeed, agents take a decision even when they decide to form a "normative belief", and then to form a new (normative) goal (in Conte et al., in press, this is called norm-*acceptance*), and not only when they decide whether to execute it or not.

As was observed by Shoham and Tennenholz (1992), computational models of autonomous norm-acceptance are lacking in the field of multi-agent systems. There have recently been several attempts to model aspects of organizational/social structures thorough notions such as social commitment, delegation/responsibility, role (Castelfranchi 1995, Singh 1997, Cavedon & Sonenberg 1998), which allow indirect ways of coping with social/collective behaviours controlled by social norms. Those notions stem from the need for avoiding or relaxing strict external constraints on agents, and from a want for cooperative, coordinated behaviour among a set of really autonomous agents. The resulting models are indeed characterized by a greater flexibility and by some local decision making mechanism.

For example, the notion of social commitments permits to translate in a subjective and local form the more general network of constraints over agents, since commitments are usually founded on cognitive ingredients, i.e. beliefs, goals, intentions, etc. Through social commitments, the agents can link their interpersonal relations to the more general norms and conventions adopted in a group or a team. A similar function has been accomplished by the concept of role, which involves goals and responsabilities, and has been used, for example, in the definition of team plans for collaborative actions (Barbuceanu 1997; Kinny et al. 1994; Tambe 1996). Finally, several studies have considered the notion of delegation, focusing on the responsibilities linked to delegation (for example, how obligations, duties, rights, and so on, between a user and its PDA – personal digital assistent – are distributed) and on definition of the delegation itself (Santos & Carmo 1996; Castelfranchi & Falcone 1998).

3. How to Represent Autonomous Normative Agents

In Section 2, we have discussed some problems and results presented by the separate treatment of norms in the two fields of legal theory and MAS. We have in

particular observed that legal studies have mainly focused on commands deliberately issued by a normative authority, while multi-agent research have mainly addressed conventions emerging from coordination processes. Moreover, formal approaches in legal studies have focused on the representation of normative modalities (and on the possible conflicts between prescriptions and behaviours), while multi-agent research has focused on the impact of norms on the agent's behaviours (and on the emergence of norms out of behaviours). However, we have also seen that both discipline have insufficiently considered the internal (mental) states linked to the norm acceptance, and the deliberative processes which produce those states.

It only by addressing this issue that we can hope to answer some major questions as the following ones:

(a) What is the relation between explicit prescriptions and social conventions? How are this notion linked to that of a legal norm?
(b) How are norms, in both senses, implemented into intelligent and autonomous agents? More specifically, what is the difference between normative and non-normative reasoning and decision-making of intelligent autonomous systems?

As for the former, it would superficially seem that norms as emerging conventions have nothing in common with norms as deliberate prescriptions. This is indeed true when conventions are merely viewed as pure regularities (uniformities) of behaviour of the concerned agents (as in the mainstream approach adopted in MAS) and when prescriptions are merely viewed as commands issued by an authority (as in the mainstream positivistic approach to legal theory). A connection between conventions and prescription emerges, on the contrary, when we pay attention to the point of view of the norm addressees, i.e., to agents' internal states and expectations. Both the emergence of conventions and the issuing of explicit prescriptions can than thus be viewed as processes through which normative standards are provided to agents. In both cases the adoption or the rejection of those standards is ultimately a choice which has to be performed by the concerned agent, on the bases of its own deliberation, by considering its own objectives, the content of the norms and the views and expectations of other agents.

This mental representation and evaluation of norms enhances agents' autonomy with regard both to conventional and prescribed norms, by allowing agents to (a) tell what is a social convention is and what is not, and therefore communicate with others about those conventions; (b) tell what a normative authority is and what it is not and therefore recognise (legitimate) prescription; (c) solve possible conflicts among norms.

Let us now move to the problem of distinguishing legal and social norms. From a perspective that duly recognises the mental role of norms and normative thinking and is so capable of linking conventions and prescriptions (as being two sources of normative representations) in the same deliberative process, there is no need

to superimpose the distinction between conventions and prescriptions to that of legal and social norms. We may indeed accept that there are legal rules which have spontaneously grown (customs-convention) and other legal rules which have been deliberately made (legislation-prescriptions). The distinction between law and social norm can instead, with Hayek (1976, 58) be viewed as the distinction between rules to which the recognized procedure of enforcement by appointed authority ought to apply and those to which it should not: for an agent a norm is a legal one if he believes that it should be enforced (or, in a different perspective, if he foresees that it will be enforced), and it is only a social if he believes that it should not be enforced (but that it should be followed).

However, the view that social norms do not imply enforceability in the sense in which legal norms do (i.e., by the appointed authority) still allows for a variety of *social* mechanisms which induce norm-following behaviour and promote autonomous acceptance. Not only legal enforcement, but also those mechanisms, which apply to both legal and social norms, assume that norms are the object of a mental representation. This can be clearly seen in the following cases: (a) spread of reputation: how can agents identify cheaters if they do not have a representation of the 'good', respectful guys? (b) Social monitoring, control: how could agents perceive, observe, and record others' behaviour with regard to the norms, if they had no mental representation of the norm itself? (c) Normative influencing: how could agents react to cheaters if they had no expectation concerning what is legal vs. non-legal, what is acceptable vs. unacceptable, what is conforming to the norm vs. violating it, etc.? (d) Rights, entitlements, etc.: how could agents be aware of their own rights and defend them if they were not associated to some normative belief? How could they pretend that their partners honour their contracts, respect reciprocity, keep to their promises, etc. if they were not supported by normative beliefs?

In conclusion, we believe that *autonomous normative agents*, which can keep into account different types of norms (conventional and prescribed ones, legal and social ones) must be endowed with mechanisms for recognising, representing, accepting norms and for solving possible conflicts among them. Only in this way those agents will be able to adopt a flexible approach towards normative standards: be aware of existing norms, be capable of violating them, be able of learning new ones, negotiate upon norms, convey them to others, control and monitor others' behaviours, influence and persuade them, etc.

Those agents must be capable of having norms as mental objects. This raises a host of practical and theoretical issues. How are normative representations possible? What type of representation do norms have? What is the role of such a representation? How does it work in the mind of an agent? What kind of connections should it have in order to affect the agent's deliberation?

4. The Papers of This Special Issue

This special issue may be considered as a first attempt to promote the cooperation and cross-fertilisation between legal studies and MAS, which we hope may provide the basis for the development of autonomous normative agents. It presents different approaches to the link between norms and agents, and, to a certain degree, shows how the combination of results from legal studies (especially AI & law) and from MAS research can allow us to address some of the open issues mentioned above.

In his paper "Autonomous Agents with Norms", Frank Dignum, addresses the question whether norms should be treated as a specific mental object in terms of a deontic logic-based approach. He suggests some good ideas about an agent architecture incorporating deontic operators in a BDI (Beliefs Desires Intentions) framework. To model generation of (some) norms, the speech acts theory is used, while deontic logic is used to model the concepts that are necessary for autonomous agents in an environment that is governed by norms. The author distinguishes three levels on which the social behaviour of an agent is determined, by individuating its different social interactions. The highest level is that of conventions; the intermediate level is the contract level (obligations and authorizations between agents that are usually created explicitly and only hold for a limited time); the lowest level is the private level (the agent makes private judgements between different obligations and/or goals and determines the actions it will take). Through deontic logic, the author not only explicitly describes those norms that can be used to implement the interactions among agents, but also both norm violations and possible reactions to such violations.

Consistent with the idea that normative prescriptions need to be explicitly related to mental attitudes, in the paper "Prescribed Mental Attitudes in Goal-Adoption and Norm-Adoption", Cristiano Castelfranchi shows how the representation of the hearer's mind in the speaker's mind is, in fact, much richer than is usually supposed (and that speech acts differ from one another as for the different mental attitudes the speaker is attempting to obtain from the hearer). While applying this point of view to normative prescriptions, the author argues that what is required by a norm is not only a given *behaviour* but also a *mental* attitude; therefore, the real task that should be faced is how to model normative minds rather mere behaviours. In the author's view important conflicts often arise not about what to do, nor about the decision to do or not to do, but about the different motivations for doing something, which are expected/requested by the speaker (or the norm "legislator"), and those that are offered by the the hearer (or the norm addressee). The concluding remark of the author is that, under any circumstance, a norm whatsoever is aimed at influencing the agent, that is to say, at changing its goals: norms should lead not only to factual conformity but to cognitive "obedience".

Commitment and flexibility in commitment have been addressed by Munindar Singh. In his paper "An Ontology for Commitment in Multiagent Systems: Towards a Unification of Normative Concepts", the author proposes a notion of

commitment that satisfies both principles from DAI and those from "spheres of control", a conceptual approach (introduced in the database community) used for structuring activities. Commitment is an important abstraction for characterizing, understanding, analyzing, and designing MASs. They also arise in distributed databases. However, traditional distributed databases, implement a highly restrictive form of commitment while their modern application requires greater organizational flexibility reflected upon more flexible forms of commitment. Singh proposes a framework called "spheres of commitment" that emphasizes the interplay between commitments and social structure. It defines operations on commitments and groups, distinguishes implicit and explicit commitments, and models social policies as higher-order commitments.

The issue of how a set of autonomous agents in a multi-agent system can be forced to act in accordance with norms is addressed by Magnus Boman in his paper "Norms in Artificial Decision Making". He proposes a solution in terms of decision theory, by implementing norms as inputs to the evaluations performed by a decision module. Hence no action that violates a norm will be suggested to any agent. The model for costraining action using norms operates on three levels of abstraction. The lowest level deals with manipulation by non-benevolent agents, with modifications of assessments as a result of sensitivity analyses, and with more or less ad hoc adoption to social norms by means of very delicate belief revision. The middle level deals with the filtering of certain actions in accordance with the risk profile of the agent. The highest level of abstraction deals with the acceptance of social norms. Since the basis of evaluation is the principle of maximization of the expected utility, the decision module does not allow agents to diminish the utility of the group that they belong to by their choice of action. This is a constructive interpretation of the principle of social rationality.

In their paper "Diagnosis and Decision Making in Normative Reasoning", Leendert van der Torre and Yao-Hua Tan present a special purpose formalism to formalize the distinction between normative diagnosis and decision theory. Following the authors' thought, the crucial distinction between the two theories is their perspective on time. Diagnosis theory reasons about incomplete knowledge and only considers the past. It distinguishes between violations and non-violations. It formalizes the hypothetical as-if reasoning of a judge or public prosecutor when he checks legal systems against legal principles. Qualitative decision theory describes how the norms influence behavior and is based on the concept of agent rationality. In contrast to diagnostic theory, a qualitative decision theory reasons about the future. The main characteristic of qualitative decision theory is that it is goal oriented reasoning, for example in planning. Moreover, authors using a preference-based deontic logic (PDL), show how deontic logic can be used as a component in normative diagnosis theory as well as qualitative decision theory.

Finally, Christen Krogh and Henning Herrestad in their paper "Hohfeld in Cyberspace and other applications of normative reasoning in agent technology" discuss when agents use norms, and the role of deontic logic with respect to: (i) the

design of agent programming languages; (ii) the design of agent communication languages. Two main claims of the authors are: (1) Formal deontic notions are useful when designing agent programming languages to be used in group work environments, because they make it easier to specify normative relationships between agents, as well as between agents and their users. (2) The theory of normative positions is useful when designing domain-specific agent communication languages, because it enables faster recovery from fraud or mistakes. Domain-specific protocols should be enhanced by making the agents agree upon which normative position should regulate their interaction before entering into a contract.

References

Alchourrón, C.E. (1991). Philosophical Foundations of Deontic Logic and the Logic of Defeasible Conditionals. In Meyer, J.-J.C. & Weiringa, R.J. (eds.) *Deontic Logic in Computer Science*, 43–84. North Holland: Amsterdam.

Alchourrón, C.E. & Bulygin, E. (1971). *Normative Systems.* Springer: Wien.

Alexy, R. (1989). *A Theory of Legal Argumentation.* Clarendon: Oxford.

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

Barbuceanu, M. (1997). Coordinating Agents by Role-Based Social Constraints and Conversation Plans. In *Proc. of AAAI'97.*

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

Castelfranchi, C. (1995). Commitments: From Individual Intentions to Group Organizations. In *Proc. of Int'l Conf. on Multi-Agent Systems* (ICMAS'95).

Castelfranchi, C. & Falcone, R. (1998). Towards a Theory of Delegation for Agent-Based Systems. *Robotics and Autonomous Systems*, Special Issue on Multi-Agent Rationality, Elsevier Editor, Vol. 24, Nos. 3–4, pp. 141–157.

Cavedon, L. & Sonenberg, L. (1998). On Social Commitments, Roles and Preferred Goals, In *Proc. of Int'l Conf. on Multi-Agent Systems (ICMAS'98).*

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.

Conte, R., Castelfranchi, C. & Dignum, F. (1998). Autonomous Norm-Acceptance, *Proceedings ATAL '98,* Paris, 4–7 July; forthcoming.

Gordon, T.F. (1995). *The Pleadings Game. An Artificial Intelligence Model of Procedural Justice.* Kluwer: Dordrecht.

Hayek, F.A. (1973). *Law Legislation and Liberty. Volume I. Rules and Order.* Routledge and Kegan Paul: London.

Hayek, F.A. (1976). *Law Legislation and Liberty. Volume II. The Mirage of Social Justice.* Routledge and Kegan Paul: London.

Herrestad, H. (1995). *Formal Theories of Rights.* Juristforbundets Forlag: Oslo.

Jones, A.J. & Pörn, I. (1985). Ideality, Subideality and Deontic Logic. *Synthèse* 275–290.

Kelsen, H. (1979). *Allgemeine Theorie der Normen.* Manz: Wien.

Kinny, D. & Georgeff, M. (1994). Commitment and Effectiveness of Situated Agents. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, IJCAI-94,* 82–88, Sydney.

Kinny, D., Ljungberg, M., Rao, A., Sonenberg, E., Tidhar, G. & Werner, E. (1994). Planned Team Activity. In Castelfranchi, C. & Werner, E. (eds.) *Artificial Social Systems, LNCS,* Vol. 830. Springer: New York.

Jennings, N. (1992). On Being Responsible. *Decentralized Artificial Intelligence* **3**: 93–102, Elsevier Science Publisher, Amsterdam.

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

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. Kaufmann: San Mateo, California.

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

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

MacCormick, N. (1978). *Legal Reasoning and Legal Theory*. Clarendon: Oxford.

Nitta, K., Shibasaki, M., Sakata, T. Yamaji, T., Xianchang, W., Ohsaki, H., Tojo, S., Kokubo, I. & Suzuki, T. (1995). New HELIC-II: A Software Tool for Legal Reasoning. In *Proceedings of the Fifth International Conference on Artificial Intelligence and Law*, 287–296. ACM Press: College Park, MA.

Pörn, I. (1977). *Action Theory and Social Science.* Reidel: Dordrecht.

Peczenik, A. (1989). *On Law and Reason*. Kluwer, Dordrecht.

Postema, G. (1982). Coordination and Convention at the Foundations of Law. *Journal of Legal Studies* 165–203.

Raz, J. (1975). *Practical Reason and Norms.* Hutchinson: London.

Raz, J. (1986). *The Morality of Freedom.* Clarendon: Oxford.

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

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

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

Singh, M.P. (1995). Multi-Agent Systems: *A Theoretical Framework for Intentions, Know-How, and Communications*, Vol. 799. Springer Verlag, LNCS.

Singh, M.P. (1997). *An Ontology for Commitments in Multi-Agent Systems: Toward a Unification of Normative Concepts*. Unpublished manuscript.

Tambe, M. (1996). Teamwork in Real World, Dynamic Environments. In *Proc. of International Conference on Multi-Agent Systems (ICMAS'96)*.

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

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