Attributing Mental Attitudes to the Normative Systems

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

Leendert van der Torre
SEN-3
CWI Amsterdam
The Netherlands
torre@cwi.nl

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

General Terms

Theory

Keywords

multiagent systems, norms, qualitative game theory

1. NORMATIVE SYSTEMS AS AGENTS

The role of norms in the formalization of multiagent systems is to stabilize the behavior of a multiagent system, and thus they play the same role for such systems as intentions do for single agent systems. However, it is still an open problem whether they should be represented explicitly, for example in a deontic logic, or they can also be represented implicitly. Boella and Lesmo [1] propose a definition of obligation in terms of beliefs, goals and desires, inspired by Goffman's game-theoretic interpretation of norms and by recursive modelling. The aim of this definition is to distinguish various reasons why agents fulfil or violate obligations.

In this paper, we extend [1]'s definition of sanction based norms by explicitly attributing beliefs, desires and goals both to the bearer of a norm and to the normative system.

Boella and Lesmo [1] propose to attribute mental states to normative systems such as legal or moral systems, as an instance of Dennett's *intentional stance* [6].

We start with a definition of Carmo and Jones. *Normative systems* are "sets of agents whose interactions can fruitfully be regarded as norm-governed; the norms prescribe how the agents ideally should and should not behave [...]. Importantly, the norms allow for the possibility that [...] violations of obligations, or of agents' rights, may occur" [7]. We restrict ourselves to the interpretation of normative *multiagent* systems as dynamic social orders. According to Castelfranchi [4], a social order is a pattern of interactions among interfering agents "such that it allows the satisfaction of the interests of some agent A". These interests can be a shared goal, a value that is good for everybody or for most of the members. E.g., the interest may be to avoid accidents.

If normative systems are dynamic social orders, the use of goals of a multiagent system can be explained by the notion of social delegation. *Social delegation* describes the behavior of a social group or institution where some of the agents, on behalf of the other ones, have to achieve some goal which is part of the plans of all members of the group or institution. We say that agents attribute the mental attitude goal to the normative system, because all or some of the agents have socially delegated goals to the normative system; these goals are the content of the obligations regulating it. The agents of the normative system thus adopt not every goal of the normative agent but only those which, if they remain unfulfilled, are considered as violations. Continuing the example, agents delegate to the normative system the goal to avoid accidents and, in order to achieve its goal, it will adopt some subgoals (such as drive on the right) which are the content of the norms it will issue to regulate the traffic. The agents will adopt these goals since they contribute to the delegated goal. However, in some cases they can decide not to adopt some norm as a goal and to violate it.

The association of violation with sanctions is explained by the notion of *social control*, "an incessant local (micro) activity of its units" [4], aimed at restoring the regularities prescribed by norms. The importance of punishment for the success of societies in evolutionary competition has been argued by Boyd *et al.* [2].

Thus, the agents attribute to the normative system, besides goals, also the ability to autonomously enforce the conformity of the agents to the norms, because a dynamic social order requires a continuous activity for ensuring that the normative system's goals are achieved. In case of sanction-based obligations, this ability is required since the application of sanctions in response to violations cannot be taken for granted: the decision of sanctioning is the result of a tradeoff of costs and advantages for the normative system. Analogously, the process of deciding whether something is a violation or not is the result of an autonomous activity.

Finally it must be noticed that in penal codes not only sanctions are explicitly associated with crimes, but also crimes themselves are defined as behaviors which are sanctioned: e.g., "murder is punished with 20 years of jail".

Thus, a normative system has the properties requested by Wooldridge and Jennings [9] for being an agent: autonomy, reactivity to changes in the environment, and pro-activeness.

AAMAS'03, July 14–18, 2003, Melbourne, Australia.

2. TOWARDS FORMALIZATION

Our framework is based on three dimensions.

1) **Agent dimension.** We distinguish between the agent A who is the bearer of the norm and the normative agent N. Further distinctions can be introduced to keep distinct the role of legislative authorities (creating norms), judicial (deciding if a behavior counts as a violation) and executive ones (applying sanctions). Moreover each type of authority may be organized in a hierarchy of levels, each one subject to the obligation to perform its task by higher levels.

2) **Mental attitudes dimension.** We distinguish between the agent's ability, its beliefs and its motivations (goals and desires). These mental attitudes are modelled by means of conditional rules in a qualitative decision theory inspired on [3]. Beliefs rules are used to infer what are the beliefs of agents using a priority relations to resolve conflicts. Goal and desires rules are used to value a decision according to which motivations remain unsatisfied. The qualitative decision theory is based on the recursive modelling of the normative agent by the bearer of the obligation: since it attributes to the normative system the status of agent, starting from agent N's beliefs, desires and goals, and its observations, the agent anticipates the decision of the normative agent, in particular, whether it considered as a violator and thus sanctioned. Then the agent bases its decision on the consequences of the normative agent's anticipated reaction.

3) **Violation dimension.** We distinguish between behavior which counts as a violation, and behavior that is sanctioned or rewarded. The notion of "counts as" is borrowed from Searle [8]'s construction of reality. Obligations are modelled as conditional rules: if the condition is satisfied but the condition is not, then agent N may decide that such situation counts as a violation and thus agent A must be sanctioned (or not rewarded). Symmetrically, permissions can be modelled as exceptional situations which do not count as violation and thus override an obligation.

The fear of sanction or desire of reward must not be considered as the only motivation to stick to an obligation: norms should be respected as such [4]. For this reason, in line with [3] we define different agent types : e.g., respectful agents try to maximize the achievement of obligations while selfish ones try to satisfy their goals; hence, they respect norms only as far as sanctions (but also the fact that they are considered violators) have an effect on their goals.

[1]'s definition of obligation is thus extended in the following way by the explicit attribution of agent A to agent N of conditional goals and desires:

1) Agent A believes that agent N wants that A does a.

2) Agent A believes that agent N desires that there is no violation $\neg V(n)$, but if N believes $\neg a$ then it has the goal to do $V(n)$: $\neg a$ counts as a violation of norm n.

3) Agent A believes that agent N desires not to sanction $\neg s$, but if $V(n)$ then it has as a goal that it sanctions agent A by doing s. Agent N only sanctions in case of violation. Moreover, agent A believes that agent N has a way to apply the sanction.

4) Agent A desires $\neg s$: it does not like the sanction.

Such a definition associates sanctions with duties since it does not presume that agent A will always fulfil its obligations. There are many reasons why agents should be able to reason about norm violation. Castelfranchi et al. [5] argue that norms can be conflicting since they are issued by different authorities and that these authorities cannot consider in advance all the possible situations where norms apply so that in some circumstances they can lead to a bad result.

Each of the three dimensions may become the basis for some violation of norms. In particular, it is possible that a selfish agent exploits the beliefs and goals of the normative agent to violate a norm without being sanctioned. As an example consider the following situations. If we consider the belief dimension, it is possible that agent A knows that agent N falsely believes that the sanction cannot be applied. Perhaps it is agent A itself who can make agent N believe that. If we consider the motivational dimension, it is possible that agent A knows that agent N has a conditional goal such that in order to fulfil it agent N has to disregard its goals to monitor and sanction violations. If agent A can make true the condition of this goal it can safely violate its obligation without the risk of being sanctioned. Finally, it is possible that agent A knows that different obligations are in conflict with each other: not only all the sanctions cannot be applied fruitfully simultaneously (in the extreme case a sentence to death makes jail irrelevant for the criminal), but also sanctions can make impossible for agent A to achieve its other duties. So agent A can take advantage from a situation where agent N cannot sanction agent A if it wants that some other more important norm is not violated.

3. CONCLUSIONS

In this paper we consider the definition of obligations in a qualitative game theory. We extend previous proposals by for example Anderson in deontic logic and Boella and Lesmo in agent theory in a qualitative decision theory extended with recursive modelling.

The definition presented here is extended to deal with hierarchical normative systems and permissions as exceptions; moreover we consider the problem of rational norm creation and of distinguishing the roles of attributed in this paper to the normative system to match Montesquieu's *trias politica*.

4. REFERENCES

[1] G. Boella and L. Lesmo. A game theoretic approach to norms. *Cognitive Science Quarterly*, 2(3-4):492–512, 2002.

[2] R. Boyd, H. Gintis, S. Bowles, and P. J. Richerson. The evolution of altruistic punishment. In *Procs. of the National Academy of Sciences*, 100:3531–3535, 2003.

[3] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the BOID architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.

[4] C. Castelfranchi. Engineering social order. In *Proceedings of ESAW00*, Berlin, 2000.

[5] C. Castelfranchi, F. Dignum, C. Jonker, and J. Treur. Deliberate normative agents: Principles and architecture. In *Intelligent agents VI*, Springer, 1999.

[6] D. Dennett. *The intentional stance*. Bradford Books/MIT Press, Cambridge (MA), 1987.

[7] A. Jones and J. Carmo. Deontic logic and contrary-to-duties. In D. Gabbay, editor, *Handbook of Philosophical Logic*, pages 203–279. Kluwer, 2001.

[8] J. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.

[9] M. J. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10(2):115–152, 1995.