

Attributing Mental Attitudes to Normative Systems

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

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

Categories and Subject Descriptors

I.2.11 [Distributed Artificial Intelligence]: Multiagent systems

General Terms

Theory

Keywords

Agent theory, normative systems, qualitative game theory

1. INTRODUCTION

In agent theory mental attitudes such as beliefs, desires, goals and intentions are attributed to autonomous computer systems to facilitate the specification, design and implementation of such systems. Using the methodology of what Dennett [4] calls intentional stance we can say that, for example, the system believes the records in its database, or that the system responds to the user's request – or it neglects it! – because it desires to do so. Boella and Lesmo [1] suggest that analogously we can attribute mental attitudes to *normative systems* like security, legal or moral systems, such that obligations of an agent can be interpreted as the desires or goals of the normative system. The motivation of their interpretation is the study of reasons why agents fulfill or violate obligations. In their definition of obligations [1] associate obligations with sanctions which provide agents with the motivation for respecting their duties.

In this paper we are interested in the attribution of mental attitudes to normative multiagent systems to facilitate their specification, design and implementation. One of the roles of obligations in multiagent systems is to stabilize the behavior of a multiagent system, and obligations thus play the same role for multiagent systems as intentions do for single agent systems. We address the following two questions:

1. How can the attribution of mental attitudes to normative multiagent systems be explained?
2. How can the attribution of mental attitudes to normative systems be used in agent theory?

2. NORMATIVE SYSTEMS AS AGENTS

Normative systems that control and regulate behavior are autonomous, they react to changes in their environment, and they are pro-active. For example, the process of deciding whether behavior counts as a violation is an autonomous activity. Since these properties have been identified as the properties of autonomous or intelligent agents [6], normative systems may be called normative agents. This goes beyond the observation that a normative system may contain agents, like a legal system contains legislators, judges and policemen, because *a normative system itself is called an agent*. Consequently, since mental attitudes can be attributed to agents, we can attribute mental attitudes to normative systems. In this section we explain the attribution of mental attitudes to normative multiagent systems by interpreting such systems as social orders.

Castelfranchi [3] defines social order as patterns of interactions among interfering agents that allow the satisfaction of the interests of agents, such as values or shared goals that are beneficial for most or all of the agents. In social orders agents attribute the mental attitude '*goal*' to the normative system, because agents *delegate* their own shared goals to the normative system, and these delegated goals become the content of the obligations regulating the system. In this way social delegation describes the behavior of a group where some of the agents, on behalf of the other ones, have to achieve a goal or obligation which is part of the plans of most or all members of the group.

For example, if agents delegate the goal to avoid accidents to the normative system, then the system may adopt the subgoal to drive on the right side of the street. This subgoal is the content of the obligation to regulate traffic. Agents adopt this goal since they contribute to the delegated goal, and they know other agents will adopt it too. However, in exceptional cases an agent may not adopt an obligation as a goal, but violate it. For example, obligations cannot all be adopted when they are conflicting, which may happen when they are issued by different authorities which cannot consider in advance all the situations in which they apply.

A consequence of this perspective on normative multiagent systems is that, due to *social control* in social orders [3], agents associate sanctions with violations. In particular, agents attribute to the normative system 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. The process of deciding whether violations are sanctioned is again an autonomous activity.

3. IMPLICATIONS FOR AGENT THEORY

There are two important advantages of the attribution of mental attitudes to normative systems.

1. Obligations can be defined in the BDI framework. The desires or goals of the normative system are the obligations of the agent. This contributes to the open problem whether norms and obligations should be represented explicitly, for example in a deontic logic, or they can also be represented implicitly.
2. The interaction between an agent and the normative system can be modeled as a game between two agents. Consequently, methods and tools used in game theory such as equilibrium analysis can be applied to normative reasoning.

For example, Boella and Lesmo [1] develop a qualitative game theory based on recursive modeling of the normative system by the bearer of the obligation: The agent bases its decision on the consequences of the normative system's anticipated reaction, using the system's beliefs, desires and goals, in particular whether the system considers its decision as a violator and thus sanctions it. Likewise recursive games can be defined from the normative system's perspective, to for example decide which norms should be created.

Moreover, our interpretation of the attribution of mental attitudes to normative systems using social order, social delegation and social control has a third advantage.

3. Behavior which counts as a violation is distinguished from behavior that is sanctioned. The normative system may autonomously decide which behavior counts as a violation, and whether violations are sanctioned.

To use the attribution of mental attitudes to the normative system in agent theory. it must be formalized. Boella and Lesmo [1] formalize the normative system using a classical decision-theoretic setting with utilities and probabilities. However, normative systems and obligations are usually not expressed using such fine-grained quantitative tools, they are instead expressed on a much more coarse-grained qualitative level. We therefore suggest a formal model of the type studied in BDI theory or in qualitative decision theory.

4. TOWARD FORMALIZATION

We envision a framework with three dimensions: the kinds of agents, the properties of agents, and the properties of norms. First, the agent A who is the bearer of the obligation must be distinguished from the normative agent N. Further distinctions can be introduced to distinguish the role of legislative (creating norms), judicial (deciding if a behavior counts as a violation) and executive authorities (applying sanctions). Moreover each type of authority may be organized in a hierarchy. Secondly, the agent's abilities, its beliefs and its goals and desires must be distinguished. For example, these mental attitudes can be modeled as conditional rules in a qualitative decision theory inspired by the *BOID* architecture [2]. Beliefs rules are used to infer the beliefs of agents using a priority relation to resolve conflicts. Goal and desires rules are used to value a decision according to which motivations remain unsatisfied. Thirdly, behavior which counts as a violation must be distinguished from behavior that is sanctioned. The notion of "counts as" is inspired by Searle's notion of the construction of reality [5].

4.1 Definition of obligation

If agent A is obliged to a , then agent N may decide that the absence of a counts as a violation of some norm n and that agent A must be sanctioned, and:

1. Agent A believes that agent N desires that A does a .
2. Agent A believes that agent N desires $\neg V(n)$, that there is no violation of norm n , but if agent N believes $\neg a$ then it has the goal $V(n)$, it counts as a violation.
3. Agent A believes that agent N desires $\neg s$, not to sanction, but if agent N decides $V(n)$ 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.

Symmetrically, permission can be modeled as an exceptional situation which does not count as a violation.

4.2 Qualitative games

We can define different agent types [2]. For example, respectful agents respect norms as such, whereas for selfish agents the only motivation to comply with obligations is the fear for sanction or the desire for reward. Intermediate agent types are possible too.

A selfish agent exploits the beliefs and goals of the normative system to violate an obligation without being sanctioned. For example, if we consider the belief dimension, agent A may know that agent N falsely believes that the sanction cannot be applied. Perhaps it is agent A itself who can make agent N believe so. 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 fulfill it agent N has to disregard its goals to monitor and sanction violations. If agent A can enable the condition of this goal, then it can safely violate its obligation without the risk of being sanctioned. Finally, agent A may know that different obligations are in conflict with each other. Sometimes sanctions cannot be applied simultaneously, e.g., in an extreme case a sentence to death makes jail irrelevant for the criminal, and sanctions can block the achievement of agent A's other duties. So agent A can take advantage from a situation where agent N will not sanction agent A, because it wants that some other more important obligation is not violated.

5. REFERENCES

- [1] G. Boella and L. Lesmo. A game theoretic approach to norms. *Cognitive Science Quarterly*, 2(3-4):492–512, 2002.
- [2] 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.
- [3] C. Castelfranchi. Engineering social order. In *Engineering Societies in the Agents World*, pages 1–18, LNAI 1972, Springer, 2000.
- [4] D. Dennett. *The intentional stance*. Bradford Books/MIT Press, Cambridge (MA), 1987.
- [5] J. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.
- [6] M. J. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10(2):115–152, 1995.