Rational Norm Creation

Attributing Mental Attitudes to Normative Systems, Part 2

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

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

1. INTRODUCTION

If a legislator introduces a new norm in a normative system, then rationality prescribes that it ensures that the norm can and will be fulfilled by agents subjected to the norm. Since agents may not follow the law, it associates sanctions with norms. But even with sanction-based obligations, some agents will look for ways to violate the norm while at the same time evading the sanction, for example by making sure that their violation will not be noticed, blocking the sanction, bribing the system, *et cetera*. Consequently, to reason about the creation of norms, we need a model of norm-evading agents. In [2] we argue that a model of norm-evading agents can be based on the attribution of mental attitudes to normative systems. In this paper we address the following two questions:

1. How can the attribution of mental attitudes to normative systems be used to reason about norm creation?

2. How can we formalize norm creation using the attribution of mental attitudes to normative systems?

2. NORMATIVE SYSTEMS AS AGENTS

Normative systems that control and regulate behavior like legal, moral or security systems 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 by [7], 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*.

The first advantage of the normative systems as agents perspective is that the interaction between an agent and the normative system which creates and controls its obligations can be modelled 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, the game theories in [1, 2] are based on *recursive modelling* 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, in particular, whether the system considers the agent a violator and thus sanctions it. Analogously, the normative system can base its decision regarding which norm to create on the consequences of the agent's anticipated reaction. There are various ways in which optimal decisions can be defined, for example by maximizing expected utility or by maximizing the set of achieved goals. When the normative system makes an optimal decision, we call it rational norm creation.

The second advantage of the normative systems as agents perspective is that, since mental attitudes can be attributed to agents, we can attribute mental attitudes to normative systems. 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 this intentional stance [4] we can say that, for example, the normative system believes that someone is guilty, or that the system sanctions someone, because it is angry.

A consequence of the second advantage is that obligations can be defined in the standard BDI framework. In particular, Boella and Lesmo [1] suggest that we can attribute mental attitudes to normative systems, such that obligations of an agent can be interpreted as the wishes, desires or goals of the normative system. The motivation of their interpretation is the study of reasons why agents fulfil or violate sanction-based obligations. The model builds on the work of Goffman [5].

Combining these two advantages, norm creation can be modelled in the BDI framework as an action with a set of pre- and postconditions, and games between the creator and the agent can be modelled as a qualitative game theory between BDI agents. Whether the agent fulfills or violates the norm depends on the agents' abilities, their mental states, their agent characteristics, and sanctions associated with the norm. *Roughly*, the normative agent can only create an obligation for p if it has a goal and desire for p, it can apply the associated sanction, and the agent desires not to be sanctioned. The consequences of the norm creation action is that the normative system is extended with a new norm, and $\neg p$ counts as a violation that will be sanctioned.

3. TOWARDS FORMALIZATION

The agent A who is the bearer of the norm must be distinguished from the normative agent N. Moreover, the role of legislative authorities (creating norms), judicial (deciding if behavior counts as a violation) and executive ones (applying sanctions) can be distinguished. The agents' abilities, their beliefs and their motivations (goals and desires) must be distinguished. For example, these mental attitudes can be modelled as conditional rules in a qualitative decision theory inspired by the *BOID* architecture [3]. Belief rules can be used to infer the beliefs of agents using a priority relation to resolve conflicts. Goal and desire rules can be used to value a decision according to which motivations remain unsatisfied.

3.1 Qualitative games

Decision problems related to norm creation run into the problem of evidence of violation. That is, agent N can only apply sanctions when it has evidence of violation, not when it has only a reason to believe that there is a violation. We may have, for example, that agent N believes that agent A will speed whenever it knows that there are no speed registrators. In such a case without speed registrators, agent N will assume that agent A will speed, but it cannot sanction agent A based on this belief. Consequently, agent N has to distinguish between the state it expects to arise, and the state it can use in its decision making.

The evidence of violation problem is related to the question *how many* levels have to be constructed in a recursive decision problem. In a decision problem, an agent only knows the initial states and considers the effects of its decisions and of the predicted effects of the decisions of the recursively modelled agents. Clearly, agent N has the decision problem which norm to create, agent N models agent A which has the problem whether to violate or not, and agent N models agent A's model of agent N whether agent A will be sanctioned or not. Moreover, and this may be less clear a priori, agent N also has a model of its own decision whether it will sanction or not. The decision whether to sanction or not cannot be based on the expected outcome, but it has to be based on observations of agent A's behavior.

3.2 Norm creation

Obligation $O_{AN}(x, s|c)$ is read as 'agent A is obliged in system N to see to it that x in context c, otherwise it is sanctioned with s', and a prohibition $F_{AN}(x, s|c)$ is read as 'agent A is forbidden in system N to see to it that x in context c, otherwise it is sanctioned with s'. Creating an obligation means that agent N adopts the desire and the goal that, if the obligation is not respected by agent A, a prosecution process is started to determine if the situation 'counts as' a violation of the obligation and that, if a violation is recognized, agent A is sanctioned. Moreover, there are several rationality constraints. For example, a precondition of norm creation is that the content of the obligation is a desire and goal of agent N, $D_N \cap G_N$, and agent N wants that agent A adopts this goal. Moreover, both agent N and A do not desire the sanction: agent N has no immediate advantage from sanctioning, while for agent A the sanction is an incentive to respect the obligation. We introduce a set of norms $NS = \{n_1, \ldots, n_m\}$, and write for the set of violation variables $\{V(n_1), \ldots, V(n_m)\}$, see [6]. Finally we write $\alpha \rightarrow \beta$ for a rule 'if α then β'.

Definition 1 (Sanction-based obligation). *Let NS be a set of norms $\{n_1, \ldots, n_m\}$ and let the variables contain the violation variables $V = \{V(n) \mid n \in NS\}$. The action of agent N to create the obligation to decide to do a with sanction s in context c, $O_{AN}(a, s|c)$, is characterized by:*

Preconditions:

1. *$c \rightarrow a \in D_N \cap G_N$: agent N desires and has as a goal that a and wants agent A to adopt a as a goal.*
2. *$c \rightarrow \neg s \in D_N$: agent N desires $\neg s$, not to sanction. This desire of the normative system expresses that it only sanctions in case of violation.*
3. *$c \rightarrow \neg s \in D_{NA}$: agent N believes that agent A has the desire for $\neg s$, which expresses that it does not like to be sanctioned.*
4. *agent N knows a way to apply the sanction.*

Postconditions:

1. *NS is extended with new norm n.*
2. *$c \wedge \neg a \rightarrow V(n) \in D_N \cap G_N$: if $\neg a$ then agent N has the goal and the desire $V(n)$: to recognize it as a violation.*
3. *$\top \rightarrow \neg V(n) \in D_N$: agent N desires that there are no violations.*
4. *$c \wedge V(n) \rightarrow s \in D_N \cap G_N$: if $V(n)$ then agent N desires and has a goal to sanction agent A.*

One way to formalize the fourth precondition is to introduce a distinction between actions and states. Then we can also distinguish between ought-to-be and ought-to-do, and between sanctions as actions or states, and violation variables as actions or states.

With multiple agents, the normative agent has the following obligation distribution problem. Given a set of goals or desires of the normative agent, how are they distributed as obligations over the agents? Typical subproblems which may be discussed are whether a group of agents can be jointly obliged to see to something, without being individually obliged. Another subproblem is whether agents can transfer their obligations to other agents.

4. REFERENCES

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

[2] G. Boella and L. van der Torre. Attributing mental attitudes to normative systems. In *Proceedings of AAMAS'03*, 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] D. Dennett. *The intentional stance*. Bradford Books/MIT Press, Cambridge (MA), 1987.

[5] E. Goffman. *Strategic interaction*. Basil Blackwell, Oxford, 1970.

[6] L. van der Torre and Y. Tan. Diagnosis and decision making in normative reasoning. *Artificial Intelligence and Law*, 7(1):51–67, 1999.

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