Fulfilling or Violating Obligations in Normative Multiagent Systems

Guido Boella
Università di Torino
Italy
E-mail: guido@di.unito.it

Leendert van der Torre
CWI Amsterdam
The Netherlands
E-mail: torre@cwi.nl

Abstract

A theory of rational decision making in normative multiagent systems has to distinguish among the many reasons why agents fulfill or violate obligations. We propose a classification of such reasons for single cognitive agent decision making in a single normative system, based on the increasing complexity of this agent. In the first class we only consider the agent's motivations, in the second class we consider also its abilities, in the third class we consider also its beliefs, and finally we consider also sensing actions to observe the environment. We sketch how the reasons can be formalized in a normative multiagent system with increasingly complex cognitive agents.

1. Introduction

There are two definitions of normative systems, one that incorporates the multiagent system and one that does not. To distinguish them we call the first one a normative multiagent system and the second one simply a normative system. Using this terminology, normative multiagent systems are "sets of agents (human or artificial) 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 actual behavior may at times deviate from the ideal, i.e., that violations of obligations, or of agents' rights, may occur" [11]. The relation between normative multiagent systems (NMAS), multiagent systems (MAS) and normative systems (NS) such as legal or moral systems may thus be described by the equation $NMAS = MAS + NS$.

A theory of rational decision making in normative multiagent systems is essential for many theories and applications in which agents are able to violate norms and such norm violations have consequences, such as theories of fraud and deception [8], threats in persuasion [12], trust and reputation, electronic commerce, virtual communities [4, 5, 13], social agents, agent-based software engineering [10], *et cetera*. This can be opposed to the Shoham and Tennenholtz' characterization of social laws in game theory [14], because their theory may be useful to study the use of norms in games and the emergence of norms in multiagent systems, but it is less useful to study the effectiveness of norms. We thus assume that norms are represented as soft constraints, which are used in detective control systems where violations can be detected (you can enter a train without a ticket, but you may be checked and sanctioned), instead of hard constraints, which are restricted to preventative control systems that are built such that violations are impossible (you cannot enter a metro station without a ticket).

There are several issues involved in a classification of the reasons why agents fulfill or violate obligations. It depends on the type of agents, because selfish agents violate more obligations than moral or social agents. Moreover, it depends on the cognitive abilities or bounded rationality of the agents. If an agent cannot reason about the possible benefits of breaking some promises, it may be better for it to build a good reputation by fulfilling as many of its promises as possible. More subtly, there is a balance between on the one hand the sanctions associated with obligations, and on the other hand the control mechanisms to enforce sanctions. Moreover, there are issues concerned with agent communication, such as threats and promises.

In this paper we propose a first classification of reasons to fulfill or violate obligations, based on an increasing complexity of the cognitive abilities of agents. We consider only games between a single agent and the normative system. We do not consider joint actions, and we do not consider communication among agents, such as threats. The examples we present can be used as benchmark examples for future generations of models of normative multiagent systems. We also sketch a formalization of such increasingly complex models.

2. Motivations

Normally an agent fulfills its obligations, because otherwise its behavior counts as a violation that is being sanctioned, and the agent dislikes sanctions. Moreover, it may not like it that its behavior counts as a violation regardless of the sanction, and it may act according to the norm regardless of whether its behavior counts as a violation, because it believes that this is fruitful for itself or for the community it belongs to. There are three categories of exceptions to this normal behavior.

First, an agent may violate an obligation when the violation is preferred to the sanction, like people who do not care about speeding tickets, or Beccaria's famous argument against the death penalty [2], which says that death as a sanction makes the sanctions associated to other crimes irrelevant. In such cases of norm violations, the system may wish to increase the sanctions associated with the norms which are violated (for speeding) or decrease the sanction of another norm (death penalty).

Secondly, the agent may have conflicting desires, goals or obligations which it considers more important. Obligations may conflict with the agent's private preferences like wishes, desires, and goals. In the case of conflict with goals the agent has committed to, the agent has to decide whether the obligation is preferred to the goal, as well as whether it should change its mind. Different types of agents can have different attitudes towards goal reconsideration: some agents prefer not to reconsider their goals, while others do [7]. In this case, a violated obligation implies a conflict between its desires or goals, and the undesired sanction. Moreover, even respectful agents, that always first try to fulfill their obligations before they consider their private preferences, do not fulfill all obligations in the case of contradictory obligations.

Thirdly, the agent may think that its behavior does not count as a violation, or that it will not be sanctioned. This may be the case when the normative system considers the cost of applying the sanction too high, but more likely it is due to an action of the agent. The agent can change the system's behavior by affecting its desires or its goals. Its action may abort the goal of the normative system to count the agent's behavior as a violation or to sanction it, as in the case of bribing, or it may trigger a conflict for the normative system. The agent can use recursive modelling to exploit desires and goals of the normative system thus modifying its motivations and inducing it not to decide for the sanction.

3. Abilities

Abilities or action repertoires introduce new possibilities to influence the normative system. In particular, there are some new twists and angles with respect to the third class of examples, where the agent can trigger conflicts for the normative system related to the normative system's abilities. For example, the agent could launch a denial of service attack against the normative system. When it is more important for the system to stop the attack than to apply the sanction and it cannot do both actions, then the system may drop its decision to sanction the agent. Note that the normative system would like to stop the attack and sanction the agent, in contrast to the bribing example in the previous section where the normative system only wants to take the bribe but no longer desires to sanction the agent. Alternatively, the agent can make it more difficult, and hence more costly, for the normative system to execute the prosecution process and the sanction, when the agent's actions 'trigger' a side effect of the action of sanctioning. For example, the agent can use proxy servers to connect to the system, so that it is more difficult for the normative system to block the agent's connections. This additional cost may be too high for the normative system enforcing the respect of the obligation. Moreover, there are two new kind of examples, either due to the inability of the agent itself, or due to the inabilities of the normative system.

Fourthly, no motivation can lead an agent to fulfill an obligation if it cannot achieve it: e.g., an agent may have already reached its quota of disk space, so it has no space left to put at disposal of the community it belongs to. Moreover, the difference with the second class of examples in the previous section is that with abilities there is not necessarily an explicit conflict in the obligations posed by the normative agent: the agent may not have enough resources to fulfill all the obligations, or the actions for fulfilling the obligations are incompatible. For example, it cannot use the disk space for installing software it needs.

Fifthly, the normative system can be unable to count the behavior as a violation or sanction it. This may again be the normal behavior, but more likely it is caused by an action of the agent, either by blocking the sanction directly, or by creating a conflict with other obligations (the latter we considered above). The agent can manipulate applicability conditions of the sanction by making it impossible for the normative system to prosecute it or to perform the sanction. For example, the normative system is not able anymore to block the agent's connections, since it has changed its IP address.

4. Beliefs

The introduction of beliefs introduces the possibility to mislead the normative system. Thus far we discussed five classes why agents violate obligations discussed thus far, i.e., they may prefer a violation, they may be in conflict, they may be unable to fulfill the obligation, they may prevent the sanction, or the sanction cannot be applied. Beliefs add another dimension to each of these classes. There can be misbelief, uncertainty or ignorance about the agent's motivations and the system's sanctions, about the agent's abilities, and, most interestingly, there can be misbelief, uncertainty or ignorance of the normative system that can be exploited. This includes nested beliefs, i.e., beliefs of the agent in a (mis)belief of the normative system.

In the first and second class of examples, when an agent prefers a violation, this may be based on a misbelief that the sanction is too low or on a misbelief that there is a conflict between the obligation and a desire or another obligation.

In the fourth class of examples, an agent may believe they it is not able to fulfill its obligations, whereas it is. Alternatively, it may be able to fulfill its obligation but it does not know how. There is a misbelief or ignorance about the agent's abilities.

In the third and fifth class of examples, and the most interesting ones, beliefs introduce new kinds of examples due to the agent's ability to let the normative agent abort its goal or due to conflicts of the normative agent. In the previous section, the prime examples of letting the normative agent abort its goal is by bribing him or influencing him by blocking the sanction or making the sanction too expensive. However, with beliefs the agent can also mislead the normative agent into making him believe that he cannot sanction the agent, whereas he could. Assume again that to apply the sanction, some precondition must be true. Rather than making this precondition false, the agent can make the normative system believe that the precondition is false. For example, the agent can make the system believe that it has changed its IP address without actually doing so. The normative system will give up trying to apply the sanction since it falsely believes that it would be just a waste of effort. Analogously, the agent can make the system believe that it is in conflict, for example it can make the system believe that there is a denial of service attack when there is no such attack. Again, the agent can exploit recursive modelling in order to foresee these reactions of the normative agent.

5. Observations

Finally, there are cases of misleading the normative system related to partial observability of the normative system. Up to this point we assumed that normative system can observe everything. But if the normative system is not immediately acquainted with the fulfillment of the obligation, then the agent has to supply proof to the system that the obligation has been fulfilled. Alternatively, if the burden of proof is with the normative system, the system always has to do sensing actions before it can counts behavior as a violation, and sanction it. This leads to a further twist in the examples discussed in the previous section: the misbelief, uncertainty or ignorance can be based on the agent influencing the sensing actions of the normative system.

Inspiration for these most complex examples can be found in crime stories. It is a well known problem of criminals that they cannot drive around in expensive cars, as this may give rise to criminal investigations. Once criminal investigations have started, one can drop fake proof like DNA traces at the crime scene to distract the investigators. Similar kinds of observations occur also with respect to all beliefs based on sensing actions. Using recursive modelling and a model that contains sensing actions, it is clear that one should not attract attention.

6. Towards formalization

To formalize the relation between multiagent systems and normative systems, Boella and Lesmo [3] propose to attribute mental states to agents as well as to normative systems, a proposal which may be seen as an instance of Dennett's *intentional stance* [9]. This approach can be contrasted to other approaches used in normative multiagent systems. For example, in deontic logic [15] one abstracts away from the norms to study logical relations among obligations. Moreover, in the BOID architecture [7] there is no set of norms and norm descriptions, but instead the agent description is adapted such that obligations (O) are added to the mental states of agents. This can be interpreted as a kind of internalization of the normative system by the agents, or as an abstraction. Alternatively, neither the normative system nor the obligations can be represented explicitly using a reduction, for example the so-called Anderson reduction [1] defines obligation of p as the necessity that the absence of p leads to a violation, $O(p) = \Box(\neg p \rightarrow V)$.

In the remainder in this paper we sketch how the four systems can be formalized based on Boella and Lesmo's idea of a normative agent, using explicit normative systems and violation predicates. In the description of these systems, we distinguish as usual between the *structure* of the system, which remains relatively stable in time, and its *behavior*. The structure of a multiagent system typically consists of a set of agents, roles, collaborations, datatypes, etc., together with descriptions of the individual agents and other concepts, such as for example their mental attitudes. Here we only discuss the structure. A multiagent systems consists of a set of agents $A = \{a_1, \ldots, a_n\}$ together with an agent description (AD) mapping agents to mental attitudes like desires (D) and goals (G), and the normative system contains a set of norms $N = \{n_1, \ldots, n_m\}$ together with a norm description (ND) mapping norms to violability conditions (V) and sanctions (S), such that:

$$\langle A, AD : A \rightarrow D \times G, N, ND : N \rightarrow V \times S\rangle$$

Boella and Lesmo's basic ontology extends the set of agents with a so-called normative agent a_{n+1} that among its set of actions has actions that counts certain behaviors as violations of norms, and that enforces sanctions on agents due to norm violations.

In the first normative system that models motivations, we do not need beliefs and goals, but we need to add the normative system to the set of agents. Moreover, to encode the decisions of the agents, we add a set of decision variables X. Finally, we add a distribution of the normative system's goals (GD) to the agents (reflecting their obligations). To resolve conflicts between desires, a priority ordering on them may be added too:

$$\langle A \cup a_{n+1}, AD : A \rightarrow X \times D \times G, N, ND, GD\rangle$$

In the second theory we add abilities, by allowing a subset of X not to be assigned to any agent, representing the environment or the external world, and we add a set of rules representing the effects (E) of the decisions of the agents and the real world. These effects are assumed to be known to all agents. In comparison to the first theory, an agent may now desire or have as a goal a parameter, which it is unable to achieve:

$$\langle A \cup a_{n+1}, AD, E, N, ND, GD\rangle$$

In the third theory we add beliefs (B) to the agent description. These beliefs can be seen as effect rules relativized to the agents, which moreover can be false. To resolve conflicts among these belief rules we may also introduce a priority ordering on the beliefs:

$$\langle A \cup a_{n+1}, AD : A \rightarrow X \times B \times D \times G, E, N, ND, GD\rangle$$

In the fourth theory we add observable propositions (OP) for each agent:

$$\langle A \cup a_{n+1}, AD : A \rightarrow X \times B \times D \times G \times OP,$$
$$E, N, ND, GD\rangle$$

In the theories, we define obligations as goals of the normative system together with additional clauses for the violability and sanctions. For the behavior of the system we introduce states which in the first theory consists of the decisions and in the other theories of the decisions together with the environment. Moreover, we define the order of the games which we consider, for which we restrict ourselves to first a decision of the agent, then a decision of the normative system.

7. Summary

In this paper we presented reasons why agents may fulfill or violate obligations in normative multiagent systems, and we classified examples according to the complexity of the cognitive abilities of the agent involved. We also sketched how the models of normative systems can be formalized.

There are two lines of research. First, formalizing the normative systems presented in this paper. The first class of examples related with motivations only has been formalized in [6]. Secondly, we intend to extend the classification to examples in which more agents are involved, including communication likes threats and promises.

References

[1] A. Anderson. A reduction of deontic logic to alethic modal logic. *Mind*, 67:100–103, 1958.

[2] C. Beccaria. *Dei delitti e delle pene*. Livorno, 1764.

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

[4] G. Boella and L. van der Torre. Local policies for the control of virtual communities. In *Procs. of IEEE/WIC Web Intelligence Conference*, pages 161–167. IEEE Press, 2003.

[5] G. Boella and L. van der Torre. Norm governed multiagent systems: The delegation of control to autonomous agents. In *Procs. of IEEE/WIC Intelligent Agent Technology Conference*, pages 329– 335. IEEE Press, 2003.

[6] G. Boella and L. van der Torre. Normative multiagent systems. In *Procs. of Trust in Agent Societies Workshop at AAMAS'04*, New York, 2004.

[7] 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.

[8] C. Castelfranchi and R. Falcone. Social trust: A cognitive approach. In C. Castelfranchi and Y. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer Academic, Dordrecht, Holland, 2002.

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

[10] N. R. Jennings. On agent-based software engineering. *Artificial Intelligence*, 117(2):277–296, 2000.

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

[12] S. Kraus, K. Sycara, and A. Evenchik. Reaching agreements through argumentation; a logical model and implementation. *Artificial Intelligence*, 104:1–69, 1998.

[13] L. Pearlman, V. Welch, I. Foster, C. Kesselman, and S. Tuecke. A community authorization service for group collaboration. In *Procs. of the IEEE 3rd International Workshop on Policies for Distributed Systems and Networks*. 2002.

[14] Y. Shoham and M. Tennenholtz. On the emergence of social conventions: Modeling, analysis and simulations. *Artificial Intelligence*, 94(1–2):139–166, 1997.

[15] G. H. von Wright. Deontic logic. *Mind*, 60:1–15, 1950.