Ubi Lex, Ibi Poena: Designing Norm Enforcement in e-Institutions

D. Grossi, H. Aldewereld, and F. Dignum

Institute of Information and Computing Sciences
Utrecht University
The Netherlands
{davide,huib,dignum}@cs.uu.nl

Abstract. The viability of the application of the e-Institution paradigm for obtaining overall desired behavior in open multiagent systems (MAS) lies in the possibility of bringing the norms of the institution to have an actual impact on the MAS. Institutional norms have to be *implemented* in the society. The paper addresses two possible views on implementing norms, the so-called *regimentation* of norms, and the *enforcement* of norms, with particular attention to this last one. Aim of the paper is to provide a theory for the understanding of the notion of enforcement and for the design of enforcement mechanisms in e-Institutions.

1 Introduction

The purpose of electronic institutions (e-Institutions) is to guarantee the overall behavior of an open multi-agent system (MAS) to exhibit desired properties without compromising the agents' autonomy. Aiming in particular at easing interactions and enhancing trust between agents [11]. This is accomplished through norms directed to the agents taking part in the society which specify the behavior that the institution expects from the agents. As such, institutions can be seen as normative systems [1], i.e., as sets of norms.

Institutions do not have access to the internal states of the agents and hence, they cannot modify them in order to avoid any incongruence between the goals of the agents and the norms of the institutions. Therefore, the problem arises of how to let those norms have an effective influence on the activities of the agents. This is the problem of *norm implementation*. This issue consists of two main aspects.

First, there is an *interpretation* issue concerning the concepts used in the formulation of the norms in terms of the ontology used at the society level. It is well-known feature of normative codifications (especially in legal systems) to be "open-textured" [6] or abstract, that is, to be in need of interpretation in order for them to be translated into norms which are meaningful for the regulated society. This is what we have called the "ontological" aspect of norm implementation [4] or, to use terminology common in legal and social theory, the "constitutive" aspect [9]. For instance, an institution might require personal data to be treated according to specific procedures. The notion of "personal data" is of an

abstract nature and, in order for the norms concerning the treatment of personal data to be implemented, a clear specification of what *counts as* personal data in the given institution should be made precise. Much attention to this issue has been dedicated by the authors in previous work (see for instance [5,4]). The present paper will leave the problem of the interpretation of norm codifications aside.
Second, there is the issue concerning the design of appropriate "enforcement mechanisms" required to push the society toward the compliance to the norms of the institution. For instance, if personal data is not treated in accordance to the institutional regulation, the institution should trigger some kind of reaction. This broad notion of "institutional reaction" corresponds to what we call here enforcement.
The present paper focuses on this last point, aiming at discussing a theory for understanding the implementation of norms in institutions and the design of enforcement mechanisms.
The core of the enforcement implementation strategy presented in this paper is summarized in the saying "Ubi lex ibi poena" ("where there is law, there is sanction"). In other words, if norms are to be enforced, then the institution should specify and handle sanctions for every possible violation of the norms. The paper is trying to give some answers around two concrete questions surrounding the enforcement: How do institutions handle violations and specify enforcement mechanisms? And how should sanctions be designed in order to be effective for the enforcement of norms in eInstitutions?
In Section 2 we discuss different enforcement strategies (regimentation vs. reaction). The effect of these different enforcement strategies on the society are discussed in Section 3. In Section 4 we discuss what are the possible sanctions that an institution can take in a society consisting of software agents. In Section 5 we give some conclusions and areas for future work.

2 Dealing with Violations

There exists an obvious way in which the compliance to the norms of an institution can be implemented, namely by making the violation of the norms impossible. When this is the case we talk about *regimentation* ([7]): norm compliance is unavoidable, and hence, with respect to what is stated by the norms of the institution, the space of the agents' autonomy is strongly limited. This typically happens in e-commerce: when shopping on the web, you cannot get your goods delivered before giving consent for using your credit card number for paying those goods.
Regimentation guarantees the compliance of the society to the norms of the institution. However, it has been argued, for instance in [2], that violations can be functional for the society as a whole. Even stronger, if no violation can occur, if nothing can go wrong, it does not make sense any more to talk about norms at all. From the agent point of view, a regimented norm, is just a fact.
With *enforcement* we mean the *reaction* that the institution specifies to respond to a violation of its norms. Enforcement presupposes, therefore,

the possibility of violation. Institutions aim at regulating the behavior of agents through norms, but it is commonplace that norms are useless if the violation of those norms is ignored (to quote the Romans again: "*ubi culpa est, ibi poena subesse debet*", that is, "where there is a violation, there must be a sanction"). In other words, the enforcement of a norm by an institution requires the institution to be in the condition of recognizing the occurrence of violations of that norm in the society and to react upon them. Not surprisingly, this *check-react* enforcement mechanism is specified by means of more norms. Enforcement is sought through further regulating the domain, i.e., adding norms imposing checks and norms specifying reactions to the occurrences of a given violation. Regulations on tax evasion are a typical example in this sense: tax payment is impossible to be regimented but checks, which could detect possible violations, are made obligatory. Once the detection takes place, precise reactions are also specified and made obligatory.
On the basis of these considerations, we can isolate three types of norms involved in the specification of institutions. In fact, the whole statute of an institution could be analyzed in terms of sets of norms of these types. There is a set of *substantive norms* which consists of those norms which describe the society's behavior desired by the institution, and there is a set of *enforcement norms* consisting of norms regulating checks and reactions on violations of other norms.
The following is an example inspired by the domain concerning the policies for data protection followed by the Spanish National Transplant Organization in the organ allocation process [10].

Example 1. (Types of norms for the specification of institutions)

Substantive norm "The National Transplant Organization is not allowed to use racial data for allocating organs to patients".
Check norm "The inspecting authority should perform random checks of the compliance to the previous norm every two months ...".
Reaction norm "If racial data are used in the allocation process, then the hospital has to be fined accordingly."

The enforcement activity can thus be split in two sub-activities: check and reaction. *Check norms* deserve some further comments. They specify the way the institution is supposed to perceive the occurrence of violations. Needless to say, this can happen in many different ways. Either directly, via random checks, like in the above example; or via constant monitoring activity, like a referee in a sport match. Or indirectly, allowing agents to denounce the occurrence of a violation and then verifying their claim. This last checking activity is of an intrinsically more complex nature, calling for the establishment of tribunal-like sub-institutions within the main institution. It would be appropriate, in this case, to talk about check *sub-institutions* rather than *check norms*. For the present paper, we leave these complexities aside focusing rather on direct forms of checks.
Via such a normatively specified enforcement of the substantive norms, the enforcement issue is just lifted up to the set of enforcement norms because, if not regimented, those norms could be violated and be thus

in need of enforcement. In principle, this pattern could be endlessly iterated unless there exists a final enforcement level, whose norms are all regimented, or whose violations are not punished (see Figure 1).

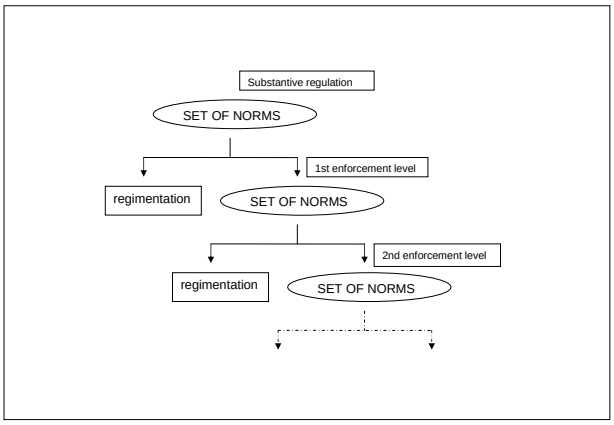

Fig. 1. Norms implementation between regimentation and enforcement.

As a matter of fact, such a cascade is precisely how real human institutions seem to be organized, where several levels of enforcement regulations may be recognized. Violations on the last level are not considered. I.e. rulings of a supreme court are supposed to be final (even though they might be violating a norm). In this sense it seems very interesting that instead of a full regimentation, the devising of a deep (i.e. structured on more enforcement levels) normative guided reaction appears to offer an efficient implementation strategy, granting at the same time a certain institutional flexibility and the room for institutional change and development. It is finally important to notice that, although we have somehow drawn a neat line between the regimentation approach and the approach leaving the possibility of violations open, an institution will most likely choose for a mixed approach deciding to regiment a (small) number of norms, and to enforce the others. We will expand on this crucial issue in the next section.

3 Different Enforcements, Different Societies

The way in which we have framed institutional implementations of norm enforcement offers a straightforward ground for showing in what precisely enforcement strategies can differ, and what kind of impact they have on societies. Consider the following (most appearing) cases:

1. A set of norms is implemented via direct regimentation;
2. A set of norms is implemented via regimentation of the set of first level enforcement norms, i.e. all occurrences of violations of the substantive regulation are sanctioned;

3. A set of norms is implemented via regimentation of the set of reaction norms of the second enforcement level, while the violation of check norms of this level is left ignored.

In Case 1 violation is impossible.In Case 2 violation is possible but the reaction is absolutely certain. This would result in creating perfect deterrents. Agents would violate the norms only if they consider the benefits obtained via violating the norm higher than the disadvantages originating by the institutional reaction.
It is clear that only in Case 3 it is possible to violate a substantive norm without any reaction on that precise violation to occur. This can happen because of a failure in complying with the corresponding check norm or with the reaction norm at the first enforcement level. If the violation of the substantive norms does not happen to be detected at the first enforcement level, then no reaction at all would follow, because at the second enforcement level only the reaction norms of the first level are regimented. This happens, for instance, when one does not get caught by the police while exceeding speed limits (they were not checking): a violation occurs which is not detected and, as a consequence, no reaction is taken. If, on the other hand, the violation of the substantive norms is detected, but still no reaction is undertaken, then the second enforcement level would automatically detect this violation occurred at the first enforcement level and react to it. This would correspond to the (unrealistic) case in which a police agent who, though detecting a violation of speed limits, does not proceed to issue a fine, is automatically sanctioned.

3.1 An example

In the previous section we sketched how institutions can implement norm enforcement over different levels of regimentation. When designing an institution this would lead of course to the question how many levels the institution should use. What are the consequences for the society when one, two, three, or more levels are used? In this section this question is elaborated by means of a simple example. We take into consideration two possible implementation strategies of an institution that two agents can use in order to play a chess match.

Example 2. (Electronic chess)
Let us first consider what happens in an electronic chess match. Players cannot move pieces other than in the way prescribed by the rules of the game, that means that they cannot violate them: the set of actions they can perform within the game is limited, and each of these actions is norm compliant. There is no possibility for them to move the rook as if it were a bishop. For these reasons electronic versions of the game of chess constitute a clear example of *regimentation* of a substantive regulation. Agents cannot do anything else than playing chess according to the rules.

It is instructive to notice that the AMELI framework [3] falls under this category. In fact, in AMELI every agent is coupled with an institutional agent, the "governor", which acts as a filter on the agent's activities letting only allowed actions to actually take place. Governors are, as

such, an excellent example of norm implementation based on the full regimentation of the set of substantive norms. It provides for a clear and protected environment. However, it is not very flexible to change (all possible moves of the game in every situation must be known beforehand).

Example 3. (Chess with flawless referee agent)
A variation on the previous example would be the use of an automatic agent referee regimenting the first enforcement level norms. Such a referee would always recognize violations and react to them. What would be the difference of this implementation of the chess institution with respect to the one described before? In that implementation, the agents could not do anything but play chess, while here they would have a wider range of actions at their disposal such as, for instance, making illegal moves on purpose in order to distract (or signal) the opponent.
The resulting games, would therefore be quite different from the one implemented in the previous example, even though the set of substantial rules (the rules of chess) is the same.

Example 4. (Chess with referee agent)
Consider now how a chess match in a standard live contest is devised. The two players are not subjected to any regimentation: there is no limitation of the set of actions available. They have the possibility to move rooks as bishops, they can thus violate the rules of chess. However, there is a further set of norms stating precisely how to react to a violation. There might for example be a third party involved, namely a referee agent, whose task is to detect violations and react to them in specific ways (or to whom suspected violations can be reported by the players). We can then think of a norm, addressed to the referee, stating that the referee ought to check what happens on the chessboard (check norm), to signal an occurring violation and to intervene in the game suspending it and ordering the faulting player to retract its move (reaction norms).
Nevertheless, this might not be enough. Violations can indeed occur also at this level and the same implementation problem is then shifted to the first enforcement level. What should happen if the referee does not detect a move that is not allowed, or does not sanction a player? A further set of norms siding, this time, the first enforcement regulation provides answers to these questions. A new enforcement level, namely a *second enforcement level*, is therefore added. This can be a contest committee which is obliged to annul a game vitiated by referee's faults and so on. As already noticed, reactive levels can in principle be added *ad infinitum*, but they are, of course, *de facto* limited. For a chess contest, two reactive levels could be reasonably enough to grant a regular chess match. However, they are not enough in an absolute sense. It is possible that the last reactive level does not behave in the expected way (reconsidering the example, suppose the committee not to annul an irregular match), at least as far as it itself is not fully regimented.
What are the new opportunities in this situation? Notice that in this situation players might violate the norms without being noticed (and sanctioned). Therefore the simple fact that a player does not violate the rules might already give him extra credit with his opponent. A notion like "trust" suddenly might become important in this game. In general,

the possible reasons for making a (illegal) move have again multiplied as well as the interpretation of them. Therefore, again, the game is enriched even though the basic rules stayed the same.

By means of this example we illustrated how different implementation strategies of the same substantive set of norms can actually give rise to radically different institutions and therefore to considerably different systems. The natural question arising is then: what would be, given a society and a set of substantive norms, the most sensible implementation strategy? And more crucially, why to allow for violations instead of choosing for a full regimentation?

3.2 eInstitutions: to Regiment or to Enforce?

The implementation of a set of (substantive) norms can be obtained either via regimentation or via the specification of an enforcement activity to be carried out by the institution. Enforcement specification takes place normatively, i.e., via adding more norms to the prior set which, thus, also require implementation. Schematically, suppose S to be the set of to-be-implemented norms, $\mathit{Regiment}(X)$ to denote the set of norms from X which are regimented, and $\mathit{Enforce}(X)$ to denote the set of norms containing X together with all the norms specifying the enforcement of X ($X \subseteq \mathit{Enforce}(X)$). The implementation of S can be formally defined as follows:

$$\mathit{Implement}(S) = \mathit{Enforce}(S \setminus \mathit{Regiment}(S)).$$

In other words, to implement a set of norms amounts to implement the set of unregimented norms together with their enforcement. This definition clearly states that the implementation of a set of norms yields a set of norms, and this is, in a nutshell, one of the main theses we are upholding here. In some sense, it is very difficult to get rid of the normative reality. The only possibility is via regimentation. In fact:

$$\text{If } \mathit{Regiment}(S) \equiv S \text{ then } \mathit{Implement}(S) \equiv \emptyset.$$

Instead:

$$\text{If } \mathit{Regiment}(S) \subset S \text{ then } \emptyset \subset \mathit{Implement}(S)$$

which means that the implementation operation should be applied again on $\mathit{Implement}(S)$.
This analysis has been led by the consideration of human institutions, but when it comes to electronic ones, some more assumptions can be made.
First of all, for human institutions it can be accepted that the violation of some norms can remain in principle ignored (see Example 4), this is not the case for e-Institutions. No designer would accept the possibility of norms the violation of which would not trigger any reaction.
Secondly, for e-Institutions, one enforcement level (level one of Figure 1) is enough. The reason is that when implementing unregimented norms,

we would expect enforcement agents explicitly programmed by the designer of the institution, and therefore we would assume them to act in perfect accordance with the principles of the institution itself[1].
Based on these considerations we can consider Example 4 as too rich (and unrealistic) in the perspective of e-Institutions. If an institution has to be designed for agents to play chess, than the possibility of an unreliable referee can be reasonably ruled out assuming that the designer of the institution would program appropriate referee agents[2]. Only two implementation choices are therefore to be considered realistic:

1. Either all substantive norms are regimented: $\mathit{Regiment}(S) \equiv S$. In this case no checking and reacting activities are necessary like in Example 2.
2. Or some (possibly all) norms are left unregimented ($\mathit{Regiment}(S) \subset S$), while what is regimented is just their enforcement like in Example 3, that is: $\mathit{Regiment}(\mathit{Enforce}(S \setminus \mathit{Regiment}(S))$.

The question amounts then to: "when is it better to choose 1 over 2 or vice versa?" In general, the preference for 2 over 1 can be dictated by two factors.
Complexity of the regimented activities Regimentation can considerably raise the complexity of the activities that agents carry out within the institution, so that for an agent to pursue its goals it would be compelled to go through unnecessarily complex procedures. This is illustrated by a simple example: consider a postal service in which the deliverer should wait for the addressee to open his/her parcels and confirm the content has been delivered in the desired state. This would rule out the possibility of deliveries of damaged parcels, but it would also make the delivery process considerably slower and inconvenient for the agents which should always be present at the delivery. In other words, regimentation can thus give rise to computationally demanding activities (see [12]) both for the institution itself, and for the agents acting within it. Formally analogous scenarios can be devised especially in the eCommerce domain, where the possibility of simple and quick transactions can be a highly desired feature.
This aspect has directly to do with the delicate balance between the two fundamental goals of e-Institutions, i.e, increase trust in agents' transactions and facilitate those transactions [11]. The point is that, although via regimentation the highest level of trust can be achieved, agents' interaction can end up being not facilitated at all.
Usefulness of the violations As we have seen in Example 3 the possibility for agents to violate the substantive regulation would allow for

[1] It is instructive to notice that this is not the case in human institutions, where the enforcement is always outsourced, in the sense that no agent can be assumed to be "programmed" by the institution: for instance, enforcement agents such as policemen do maintain private goals and believes completely inaccessible from an institutional perspective. This is why, in human institutions, the nesting of many more that just one enforcement levels is the rule.

[2] These are of course contingent assumptions on the actual state of the art in MAS and e-Institutions. Future developments in these fields would make them become possibly obsolete. It can be indeed thought of e-Institutions delegating the enforcement activity to agents of different e-Institutions.

activities which would otherwise be impossible. The agent can choose to violate the regulation and possibly incur in a sanction in order to pursue some specific goals. In Example 3 agents playing chess in an institution with a flawless referee would actually have the possibility to use a wider variety of strategies for winning the game by trying to distract the opponent via performing invalid moves. Alternatively, suppose a reputation value to be attached to each chess-playing agent so that the less often they violate the norms the higher reputation they get. In this case, the possibility to violate the norms enables also the possibility to introduce a reputation value system which might be useful for further purposes: for instance, a high reputation value might be required to access chess tournaments.
At the end, allowing for violations results in a higher flexibility of the e-Institutions which might happen to serve more purposes than the one for which it was designed. This can be a desirable feature especially in domains where more e-Institutions operate on the same society.

4 Sanctions in e-Institutions

When using an enforcement mechanism to implement norms in an e-Institution, as argued in the previous section, sanctions need to be specified to define the institution's reaction to the violations of the norms. Violations that do not trigger any reaction have no sensible meaning in an e-Institution that uses the norms and sanctions, the reaction to the violation of the norms, to direct and control the behavior of the agents participating in the institution.
In previous literature (cf. [11,8]) several kinds of sanctions have been proposed, mostly influenced by sanctions used in human institutions. Some of the sanctions involve, e.g. bans, dismissal, reputation or trust influences, fines to the agent or its owner, etc. However, when designing an e-Institution not all human sanctions make sense, like, for instance, incarceration, which is a common sanction for humans, but no direct electronic equivalent of this sanction appears useful. In general there are two ways of sanctioning agents which make sense: 1) limiting the future actions of the agents, or 2) executing an action on behalf of the agents. The first option includes, but is not limited to, sanctions such as bans and fines that are meant to restrict the agent in doing actions that are needed for it to achieve its goals (the money spent on the fine was actually meant to buy goods in an auction; the ban prevented the agent from making a bid before the auction closed). The second kind of sanctions are those where the institution changes some information (resource) pertaining to the agent which usually can only be changed by the agent itself. This might consist in changing the reputation of the agent or in paying bills on behalf of the agent, because either the agent has granted the institution this power upon entering (by signing a contract that states that the institution has the authority to issue payments on behalf of the agent in case of violations), or because the agent had to pay a deposit when it entered the institution (the deposit is then used to pay the bills and any fines that might arise).

Whatever type of sanction is chosen they are there to serve a purpose. In the following we examine the purpose of sanctions. We look at what sanctions are supposed to do and how the complexity of the agents in the institution can influence the choice of sanctions.

4.1 A Taxonomy of Sanctions

Sanctions serve different purposes in different institutions. However, there is a general purpose to sanctions that holds for all institutional environments: sanctions are there to discourage agents from taking actions that are considered unwanted or illegal by the institution. Sanctions can be seen as a deterrence, making agents less keen about performing these unwanted and illegal actions. To achieve this discouraging effect on the agents in the institution, sanctions are designed to limit the future actions of agents. For instance, fines influence the possibilities of the agent, since they make it harder for the agent to get the items it requires as the agent has less money to spend (which, of course, only really restricts the agent if it had a limited budget and the agent's owner ordered the agent to obtain lots of items). Similarly, reputation changes might limit the actions of an agent as it might influence the outcome of future negotiations and interactions of the agent.

Next to their discouraging effect, sanctions might also be used as a compensation to those most effected by the violation of the norm. In order to provide some satisfaction or compensation to those harmed by the violation, the violating agent is sanctioned. For instance, an agent might become obliged, after violating a norm, to pay an amount of money to the affected agent(s) as compensation. This difference between using sanctions as a deterrence and as a compensation signifies a difference in the role of the institution when applying the sanction. Sanctions that are solely used as a discouragement are sanctions that are applied by the institution itself, and therefore benefit the institution itself (the fines are payed to the institution, bans are applied solely to maintain order in the system). When sanctions are applied to provide a compensation to those harmed (note that the sanction will also retain its deterring nature), the institution becomes a mediator instead, interacting between the agent who committed the violation and the rest of the society.

Another difference in sanctions, as mentioned in [12], is whether the sanction is of direct or indirect nature. Direct sanctions are those that influence the agent immediately and are noticeable directly. These include fines, bans and other "corporeal" sanctions. Indirect sanctions, on the other hand, influence the agent on a kind of meta-level, such as reputation changes or trust related sanctions. Those sanctions might not be noticeable immediately but can influence the agent for a longer period of time. Combinations of both types of sanctions can be used as well.

The choice between using a sanction merely as deterrence or adding a compensational value to it depends on the norm and domain in question. If the violation of the norm harms other agents, and these 'victims' require support to overcome this harm, a compensation might seem appropriate. However, if the norm only affects the institution, no compensation is needed. Similarly, the choice between the usage of direct and

indirect sanctions is entirely up to the domain and norm in question. If indirect sanctions have an equal deterring value as direct sanctions, indirect sanctions can be used just as well. In a domain, however, where reputation plays no role, an indirect sanction (in this case, lowering the agents reputation value) has no value and a direct sanction should be used instead.

4.2 Sanctions and Types of Agents

Whatever purpose the sanctions might serve in a certain institution, the complexity of the agents in the system must be understood to determine the effectiveness of the sanction. A system that is trying to discourage agents from violating the norms by applying bans might be quite successful when the agents in the system feel bad about being banned, or are unable to complete their goals because of the ban. However, if the agents do not mind the ban the sanction fails its purpose.
The hierarchy of types of agents' autonomy developed in [13] can be used to distinguish, for each level of autonomy, what the impact of sanctions can be and which sanctions are suitable for the cognitive structure of the agent. The hierarchy of [13] distinguishes the following types of autonomy in agents (also see figure 2):

Type I Reactive Agents: Agents whose autonomy completely resides in the combination of environmental cues and system properties.

Type II Plan Autonomous Agents: Agents that are autonomous in their choosing the sequences of actions (plans) to obtain goals. The goals itself are either inherent to the agent or triggered by requests from other agents.

Type III Goal Autonomous Agents: Agents that are autonomous in making decisions about goals (which have become their interests), enabling them to choose their "prevailing interest", considering its goals. It determines which states of the world are desired, given the goal satisfaction and its goal priority.

Type IV Norm Autonomous Agents: Agents with the capabilities to choose goals that are legitimate to pursue, based on the norms of the system. Moreover, norm autonomous agents are equipped to judge the legitimacy of its own and other agents' goals.

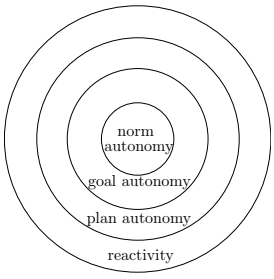

Fig. 2. A Taxonomy of Autonomy

The lower level agents, i.e. types I to III, have no idea of a sanction (they have no conception of what a sanction is). To these agents, a sanction applied by the system is nothing more than an environmental reaction to the situation at hand (or to the action they have just performed). This makes informing the agents about the norms a bit harder, as the norms need to be translated to situational causal effects that are triggered by actions in various situations. The sanctions become a necessary causal effect of the actions prohibited by the norms. However, directing and controlling the agents is a bit easier for the lower types of agents, as punishing agents by making them unable to reach their goal is easy for agents of types I and II. These agents can easily be prevented to achieve their goals by making them unable to do an action (making the sanction not as much a punishment for the agent, but more an incentive for the developer to redesign the agent to become norm-compliant). This is a bit harder for agents of type III , as these agents are more capable of coming up with alternative ways to achieve their goal (or can pursue alternative goals, making the punishment less effective).

Type IV agents are even a bigger problem, since they have a clear conception about what a sanction is and when a sanction will be applied. These agents can reason about the results of their actions in a normative manner, i.e. they take the norms into account to determine if an action in a certain situation is acceptable or if it will trigger a violation. This means that if a type IV agent violates a norm, the agent has probably reasoned that violating the norm is the only or the most efficient way to achieve it's goal, and a punishment is therefore only an increase in cost for the agent doing the action (while this increased cost has been fully taken into account in the decision of the agent). Moreover, since agents of type IV have the same capabilities as agents of type III, the sanctions loose even more of their deterring effectiveness.

A big problem, however, is that no guarantees can be given whether the sanction has the right effect on all the agents possibly joining the institution. To design sanctions to work for agents, assumptions about the inner working of the agents have to be made; what effect will the sanction have on them? Will they replan and try again, or will the sanction make them sorry about what they did?

In human institutions, such assumptions about the inner process of humans can be made, and such assumptions are correct most of the time (we know how most of us think, react to certain stimuli etc.). Sanctions applied in human institutions are based on these assumptions to work as an effective deterrent, as humans tend to dislike spending time in prison or paying fines applied after violating a norm. Even alternative punishments, such as being put under probation, which can be seen merely as a warning, work for humans, as they apply to the moral sense of the perpetrator. For agents, however, this kind of reaction is not assured. Agents are programmed by different developers, making them heterogeneous in nature. This heterogeneity also means that the inner workings of agents can be very different between agents. Since one cannot assume that all agents work in a similar manner or have the same beliefs in certain situations, it makes designing sanctions that are really punishments for all agents very hard. Using, for instance, probations in agent environments

makes no sense, since most agents will not consider this sanction to be a warning.
If, however, one can assume that (the majority of) Type IV agents are programmed in such a manner that they will try to be norm-compliant, sanctioning these agents becomes a bit easier as the sanction is no longer seen by these agents just as a necessary causal effect to a prohibited action but as something undesirable in itself. This would mean that a norm breaking action is just less preferred by such agents than other norm-compliant actions (even if the norm-compliant action is more costly) because of the agent's desire to be norm-compliant. Sanctions can in this case rely on an intrinsic deterrence effect allowing for the specification of less drastic institutional reactions to violations (for instance fines instead of bans). However, if the willingness of agents to be norm-compliant cannot be guaranteed, the normative awareness with which Type IV agents are endowed cannot be exploited and they will have to be sanctioned in the same way as Type III agents.

5 Conclusions

In this paper we have explored two related problems that have to be solved when implementing norms in e-Institutions. First is the decision between enforcement of norms through regimentation or through reaction. An interesting first observation is that implementing norms actually implies adding more norms (albeit of a slightly different nature). Of course all conceivable levels of enforcement norms are possible. However, we have seen that in most situations the best is to have one level of enforcement norms in e-Institutions due to the fact that enforcing agents are centrally controlled (and programmed).
The second question addressed in this paper was which sanctions are useful as reaction to violations. We have shown that, although many mechanisms are based on human society, not all human-based sanctions make sense in an e-Institution. A first classification of different types of sanctions is given, but many issues still remain open. One of the first issues to be addressed is how to choose the most effective sanction from an institutional point of view. This would both deter agents from violating norms too easily, but also facilitate normal transactions between agents as much as possible.

Acknowledgments
Authors would like to thank the anonymous reviewers of COIN'06 for their useful comments.

References

1. C. E. Alchourrón and E. Bulygin. *Normative Systems.* Springer Verlag, Wien, 1986.

2. C. Castelfranchi. Formalizing the informal?: Dynamic social order, botton-up social control, and spontaneous normative relations. *Journal of Applied Logic*, 1(1-2):47–92, 2004.
3. M. Esteva, J.A. Rodríguez-Aguilar, B. Rosell, and J.L. Arcos. Ameli: An agent-based middleware for electronic institutions. In *Third International Joint Conference on Autonomous Agents and Multiagent Systems*, New York, US, July 2004.
4. D. Grossi, H. Aldewereld, J. Vázquez-Salceda, and F. Dignum. Ontological aspects of the implementation of norms in agent-based electronic institutions. In *Proceedings of NorMAS'05, Symposium on normative multi-agent systems.*, pages 104–116, Hatfield, England, April 2005. AISB.
5. D. Grossi, F. Dignum, and J-J. Ch. Meyer. Contextual terminologies. In *Proceedings of CLIMA VI*, LNAI. Springer, 2006.
6. H. L. A. Hart. *The Concept of Law.* Clarendon Press, Oxford, 1961.
7. A. J. I. Jones and M. Sergot. On the characterization of law and computer systems. *Deontic Logic in Computer Science*, pages 275–307, 1993.
8. P. Pasquier, R. A. Flores, and B Chaib-draa. Modelling flexible social commitments and their enforcement. In M.-P. Gleizes, A. Omicini, and F. Zambonelli, editors, *Proceedings of the Fifth International Workshop Engineering Societies in the Agents World (ESAW'04)*, volume 3451 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 153–165. Springer-Verlag, 2004.
9. J. Searle. *The Construction of Social Reality.* Free Press, 1995.
10. Ley 30/1979, de 27 de octubre, sobre extracción y transplante de órganos. Boletín Oficial del Estado 266, 29th april 1986.
11. J. Vázquez-Salceda. *The role of Norms and Electronic Institutions in Multi-Agent Systems.* Birkhuser Verlag AG, 2004.
12. J. Vázquez-Salceda, H. Aldewereld, and F. Dignum. Norms in multiagent systems: from theory to practice. *International Journal of Computer Systems Science & Engineering*, 20(4):95–114, 2004.
13. H. Verhagen. *Norm Autonomous Agents.* PhD thesis, The Royal Institute of Technology and Stockholm University, 2000.