

Relationships between obligations and actions in the context of institutional agents, human agents or software agents

Robert Demolombe¹

Institut de Recherche en Informatique de Toulouse
France
`robert.demolombe@orange.fr`

Abstract. The paper presents a logical framework for the representation of interactions between institutional agents, human agents and software agents. A case study is used to analyze how obligations on institutional agents are "propagated" to human and software agents, and how actions performed by these agents count as actions that satisfy the obligations imposed to institutional agents. It is shown that the relationship between the different kinds of obligations and actions can be represented in terms of the concept of "count as" proposed by Searle, of role and of causality. The logical framework focus on those three concepts.

1 Introduction

The interactions between companies and people have been analyzed from a legal point for a long time. More recently their interactions have involved computers, networks and complex pieces of software. For instance, new applications have been developed in the field of electronic commerce where individuals can buy or sale goods to other individuals or to companies. Electronic tools can also be used to organize auctions or to negotiate contracts (see [15, 12, 27]).

This new situation raises basic questions about the legal status of these interactions.

Surprisingly there are quite few norms to regulate them and a possible explanation for this situation may be that some basic concepts, like the legal status of the pieces of software that are used for these interactions, have no sufficiently clear definitions to allow lawyers to define appropriate new norms.

For example, there are research proposals where these pieces of software are defined as "normative agents" (see, for example, [35]) and it is not clear whether this terminology is used just as a metaphor for human beings or, strictly speaking, as legal entities. To make more explicit the risk of confusion we can consider the terminology used in the field of highways regulation, when a norm says, for instance, that "car speed is limited to 130 km/h". The risk of confusion comes from the fact that this norm can be understood as a norm which applies to cars or to their drivers. Nevertheless, for most people it is clear that this norm

applies to drivers. However, when we speak about "normative agents" it is not so easy to understand who are the human beings who are the counterparts of the drivers.

The purpose of this paper is to try to clarify these kinds of issues when we consider, from a normative point of view, interacting agents which can be either companies, human beings or pieces of software.

Our approach is to start from the detailed analysis of a case study (section 2) in order to exhibit the key concepts. Then, we propose a logical framework to define clear relationships between the norms that apply to agents and between the actions they perform (section 3). This framework is applied to the case study to evaluate its appropriateness (section 4). Comparisons with other works are provided in section 5 and in the last section are presented our conclusions and directions for possible further works.

2 A case study

Before to present the case study we define some terms that will be used along this paper. We call "institution" a set of norms in the sense proposed by Searle in [29] For example, the set of norms about international commerce is an institution. Following the same terminology we call "institutional agent" an agent whose existence and legal status is defined in the context of an institution. For example, a company, an hospital, a university or a foundation, where the status defines, for instance, the kind of taxes they have to pay and the legal responsibility of some of their managers.

Let's consider now a case study where it is obligatory that a company (the institutional agent I_1) pays a certain amount of money x (this action is called α) to another company (the institutional agent I_2).

This obligation holds in the context of an institution (institution T). For instance, this institution could be the set of norms about international commerce. Since an institutional agent is an abstract entity it cannot do concrete actions. However, there is a human agent (agent H_1) who holds a role (role r_1) in the organization of I_1 and there is a norm in the definition of r_1 which says that the fact that it is obligatory for I_1 to pay I_2 counts as the fact that it is obligatory for H_1 to send a check of amount x to the mail box of I_2 (action β).

The definition of the role r_1 and of this particular norm are part of the institution S which defines the organization of I_1 .

This norm defines how the obligation to I_1 "propagates" to an obligation to H_1 and this norm of S is justified, provided there is another norm in T which says that the fact H_1 has done β , acting as r_1 role holder, counts as the fact that I_1 has done α .

To fulfill the obligation to do β the agent H_1 can either do β himself or order (action ω) to another human agent (agent H_2) who holds some particular role (role r_2) to send a check of amount x to the mailbox of I_2 (action γ) (see figure 1).

Let's assume that H_1 selects the second choice. This is possible provided the following properties hold :

- 1 H_1 has in S the institutional power to order H_2 to do γ , that is, in S the fact that H_1 has done ω , acting as r_1 role holder counts as the fact that it is obligatory for H_2 to do γ , acting as r_2 role holder.
- 2 The fact that H_2 as done γ , acting as r_2 role holder, counts in S as the fact that H_1 has done β , acting as r_1 role holder.
- 3 The fact that H_2 as done γ , acting as r_2 role holder, counts in T as the fact that H_1 has done β , acting as r_1 role holder.

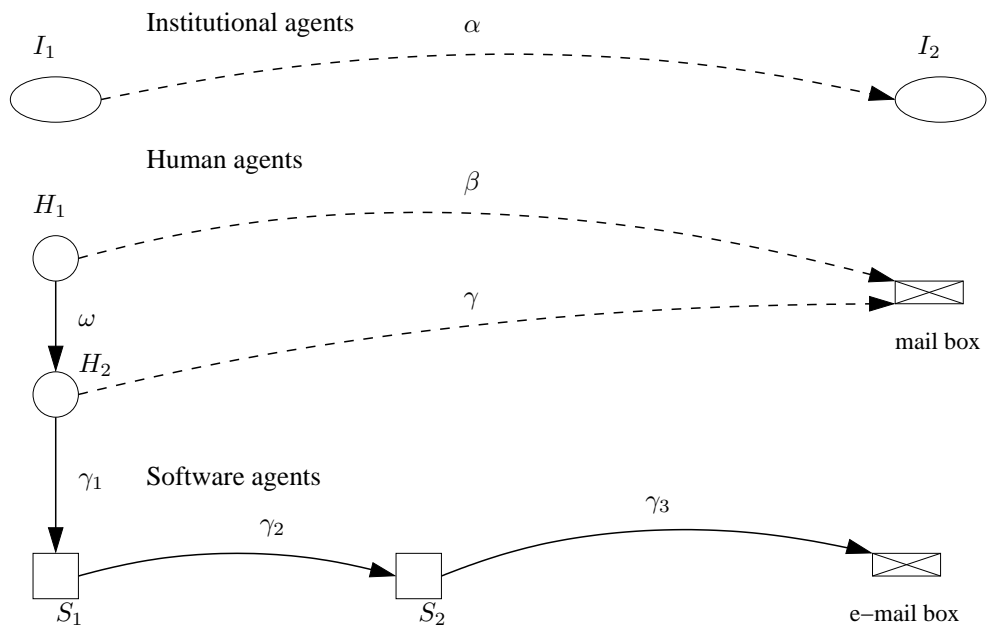


Fig. 1. Relationships between actions.

Then, in this context, H_1 does ω , that is, he orders to H_2 to do γ , and, by doing ω , H_1 "propagates" to H_2 the obligation to do β . To fulfill the obligation to do γ , H_2 can either do γ himself or use software agents. Let's assume that H_2 selects the second choice, that is, H_2 runs a software agent (agent S_1), by doing some action like filing some fields in his computer screen or clicking on some button (action γ_1) (see figure 1 where the concrete actions are not represented by dot lines).

Let's assume that action γ_1 causes the running of S_1 (action γ_2) and that the action γ_2 causes the running of another software agent (agent S_2) and that

the running of S_2 (action γ_3) has the effect to store in the electronic mailbox of I_2 a payment order of amount x .

That is possible, provided the following properties hold :

- 4 H_2 has the ability to do γ_1 .
- 5 The fact that H_2 as done γ_1 , acting as r_2 role holder, and S_1 has done γ_2 and S_2 has done γ_3 , counts in S as the fact that H_2 has done γ , acting as r_2 role holder.
- 6 The fact that H_2 as done γ_1 , acting as r_2 role holder, and S_1 has done γ_2 and S_2 has done γ_3 , counts in T as the fact that H_2 has done γ , acting as r_2 role holder.

It is worth noting that the action γ_2 done by S_1 is not determined by the propagation of an obligation but by a causal relationship between the action γ_1 done by H_1 and the action γ_2 done by S_1 . In the same way the relationship between γ_2 done by S_1 and γ_3 done by S_2 is a causal relationship.

Now, we can check that the fact that H_2 has done γ_1 guarantees that in the context of institution T the obligation for I_1 to do α is fulfilled.

Indeed, from property 6, the sequence of actions γ_1 , γ_2 and γ_3 counts in T as the action γ , and from property 3 the action γ counts as β , and it has been assumed that there is a norm in T which says that β counts as α . Also, from property 5, we can check that H_2 has fulfilled his obligations in S to do γ and from property 2, we can check that H_1 has fulfilled his obligations in S to do β .

This case study shows that actions performed by agents can be influenced or determined in different ways. Actions performed by institutional agents and human agents are influenced by obligations, that is by social laws, while actions performed by software agents are determined by causality. That is by laws of nature.

Also, the means that generates the "propagation" of these influences differ depending on the different kinds of agents. From institutional agents to human agents, the influence is propagated by norms of the kind of "counts as" (for instance, the obligation for I_1 to do α counts as the obligation for H_1 to do β).

From human agents to human agents, the influence is propagated by orders (for instance, the order ω creates the obligation to H_2 to do γ).

From human agents to software agents, the influence is propagated by actions that cause other actions (for instance, the action γ , done by the human agent H_1 causes the action γ_2 done by the software agent S_1).

From software agents to software agents, the influence is propagated by actions that cause other actions (for instance, the action γ_2 , done by the software agent S_1 causes the action γ_3 done by the software agent S_2).

The fact that actions can be "delegated" from institutional agents to human agents and to software agents is justified by norms of the kind "counts as". These "counts as" norms are themselves justified by the fact that human agents hold some given roles (for instance, action β counts as action α because H_2 holds the role r_2). The justification of the fact that actions performed by software agents count as actions performed by human agents is different. These software agent

actions have been caused by some human agent (for instance, actions γ_2 and γ_3 count as action γ because they have been caused directly, or indirectly, by the action γ , done by the human agent H_2).

The conclusion of this analysis is that to define a logical framework to represent these kinds of examples the main concepts we have to formalize are: **counts as**, **obligation**, **causality** and **role**. Other concepts may be needed like **knowledge** and **belief**. For example, to take the decision to do action β the human agent H_1 has to know that he is obliged to do this action, and to take the decision to do action γ_1 the human agent H_2 has to know the effects of this action. However, these concepts do not raise issues which are specific to the kind of case study we have analyzed and that is the reason why they are ignored in this paper.

3 Towards a logical formalization

3.1 Counts as

For the formalization of the counts as operator we have adopted the logical framework proposed in the seminal work of Jones and Sergot [20] and we have only changed minor technical details. In particular, the formalization which is presented here is only defined in the semantics.

In the following it is assumed that $\phi, \psi, \theta \dots$ denote formulas of a language of a modal propositional logic LP.

We have adopted the following notations.

$\phi \Rightarrow_S \psi$ can be read: ϕ counts as ψ in the institution S .

$D_S \phi$ can be read: ϕ is necessarily true in the institution S or ϕ is recognized by the institution S .

To define the semantics of these two operators a model M is defined as a tuple $M = \langle W, f_S, d_S, P \rangle$, where W is a set of possible worlds, P is a function which assigns a set of worlds to each atomic formula and which is extended as usual to logical connectives, f_S is a function which assigns in each world a set of sets of worlds to a set of worlds and d_S is a function which assigns to each world a set of worlds. More formally we have:

$$f_S : W \times 2^W \rightarrow 2^{2^W}$$

$$d_S : W \rightarrow 2^W$$

The satisfiability conditions¹ for these operators are:

$$M, w \models \phi \Rightarrow_S \psi \text{ iff } \|\psi\|_M \in f_S(w, \|\phi\|_M)$$

$$M, w \models D_S \phi \text{ iff } d_S(w) \subseteq \|\phi\|_M$$

where $\|\phi\|_M$ denotes the set of worlds : $\{w : M, w \models \phi\}$.

Notice that it follows from the second satisfiability condition that D_S is a normal modal operator.

¹ See [4] for the satisfiability conditions of normal modal operators and classical modal operators.

First, we have an inference rule to express the substitutivity of logically equivalent formulas in the antecedent and in the consequent of the counts as operator.

(EQV) If $\models \phi \leftrightarrow \phi'$ and $\models \psi \leftrightarrow \psi'$, then $\models (\phi \Rightarrow_S \psi) \rightarrow (\phi' \Rightarrow_S \psi')$.

Then we have two schemas to express the closure of the consequent and of the antecedent of the counts as operator.

(CC) $\models ((\phi \Rightarrow_S \psi) \wedge (\phi \Rightarrow_S \psi')) \rightarrow (\phi \Rightarrow_S (\psi \wedge \psi'))$

(CA) $\models ((\phi \Rightarrow_S \psi) \wedge (\phi' \Rightarrow_S \psi)) \rightarrow ((\phi \vee \phi') \Rightarrow_S \psi)$

We have also accepted the following transitivity schema.

(S) $\models ((\phi \Rightarrow_S \psi) \wedge (\psi \Rightarrow_S \theta)) \rightarrow (\phi \Rightarrow_S \theta)$

The links between \Rightarrow_S and D_S are expressed by the following schemas:

(D) $\models (\phi \Rightarrow_S \psi) \rightarrow D_S(\phi \rightarrow \psi)$

(C) $\models (\phi \Rightarrow_S \psi) \rightarrow (\phi \rightarrow D_S\psi)$

The intuition of (D) is that the counts as operator entails in the institution S a constraint represented by the material implication, and the intuition of (C) is that if a fact represented by ϕ counts as another fact in S , then this fact is recognized by the institution S as an institutional fact. Then, (C) allows to derive an institutional fact, represented by $D_S\psi$, from a brute fact, represented by ψ .

Notice that from (D) to (C) we can infer: $\models (\phi \Rightarrow_S \psi) \rightarrow (\phi \rightarrow D_S\psi)$

The constraints on the functions f_S and d_S which validate the above principles can be found in [20].

3.2 Obligation

Since we are mainly interested in checking whether obligations are fulfilled or violated by actions performed by agents, we have accepted the formalization of Standard Deontic Logic (see [4]) plus additional properties that link obligations and the D_S operator.

We have adopted the following notation.

$Obg_S\phi$: in the context of the institution S it is obligatory that ϕ .

In a model M for each operator Obg_S we have a function O_S which assigns to each world a set of worlds. We have :

$O_S : W \rightarrow 2^W$

The satisfiability condition of these operators is defined by:

$M, w \models Obg_S\phi$ iff $O_S(w) \subseteq \|\phi\|_M$

We have also accepted the following properties:

(R) $\models Obg_S\phi \leftrightarrow D_S Obg_S\phi$

(OD+) $\models Obg_S\phi \rightarrow (\phi \rightarrow D_S\phi)$

(OD-) $\models Obg_S\phi \rightarrow (\neg\phi \rightarrow D_S\neg\phi)$

Fulfillment and violation of an obligation are defined as follows.

Obligation fulfillment : The obligation $Obg_S\phi$ is fulfilled iff we have $D_S\phi$.

Obligation violation : The obligation $Obg_S\phi$ is violated iff we have $D_S\neg\phi$.

From (OD+) if we have ϕ the obligation $Obg_S\phi$ is fulfilled, and from (OD-) if we have $\neg\phi$ the obligation $Obg_S\phi$ is violated.

3.3 Causality

We have to formalize the fact that the performance of an action which has caused the performance of another action counts as the performance of a third action. The "counts as" operator refers to actions which have been done, not to actions which are going to be done. However, to formalize causality it is usual (see [34]) to consider the states of affairs before and after the performance of an action. Then, we need formal definitions of operators that express that an agent is going to do some action and also of operators that express that an agent has done some action. For that purpose we have defined the following operators whose intuitive meaning is presented below.

$Does_{agt:\alpha}\phi$: agent agt is going to do the action α and after performance of α the proposition ϕ holds.

$Done_{agt:\alpha}\phi$: agent agt has performed the action α and the proposition ϕ holds.

To avoid to have too many similar action operators we shall use the notation agt to denote either an agent i or an agent i acting as holder of the role r (this is represented by $i : r$). For instance, in the operator $Does_{agt:\alpha}$, if $agt = i$ we have $Does_{i:\alpha}$ and if $agt = i : r$ we have $Does_{i:r:\alpha}$, where i is acting as r role holder.

We shall use the notation $Done_{agt:\alpha}$ to express the fact that action α has been done, without reference to the effects of the action. Then, this notation can be seen as an abbreviation for $Done_{i:r:\alpha}\top$, where \top refers to any tautology.

These operators allow us to express that the proposition ϕ is true after performance of action α . To express that action α has caused the truth of ϕ we need additional operators. These operators and their intuitive meanings are presented below.

$E_{agt:\alpha}\phi$: agent agt has brought it about that ϕ holds by doing the action α .

$E_{agt:\alpha}^+\phi$: agent agt is going to bring it about that ϕ holds by doing the action α .

To define the semantics of these operators we have defined two families of accessibility relations. They are intuitively presented below.

$wD_{agt:\alpha}w'$: agent agt has started to do the action α in w and possibly other actions, and he has ended the action α in w' .

$wD_{agt:\neg\alpha}w''$: agent agt has started to do in w the same actions as he did in w' such that $wD_{agt:\alpha}w'$, except the action α , and he has ended these actions in w'' .

The relation $D_{agt:\neg\alpha}$ is intended to express the counterfactual condition with respect to the relation $D_{agt:\alpha}$ in the same way as Pörn expresses (see [25] chapter 1, section 5) the counterfactual condition with the relation D' with respect to the relation D ³. In general, there are many different ways to do α and then there are many different worlds like w' . If in some world w' the set of actions performed by agt is $\{\alpha\} \cup A$, in the corresponding world w'' such that $wD_{agt:\neg\alpha}w''$ it is

² If in w' other agents have performed some actions, it is assumed that in w'' they have performed the same actions as in w' .

³ The meaning of the counterfactual condition is defined by Pörn as: "but for i 's action it would not be the case that ϕ ".

assumed that the set of actions performed by agt is A . Then, a more formal (and more heavy) notation would require a relation with 3 arguments of the kind: $w, w' D_{agt:\neg\alpha} w''$.

A significant difference with Pörn's definition is that we have made explicit the worlds where we are before to do the action and after to do the action, while in Pörn's semantics the worlds should be interpreted as histories⁴. Other differences are the fact that agent agt may be acting as holder of a role and the fact that Pörn does not explicitly mention the name of the action α .

The satisfiability conditions for the operators of the kind *Does* and E^+ are defined in the same way whatever the agent is acting as a role holder or not. We have:

$$\begin{aligned}
& M, w \models \text{Does}_{agt:\alpha} \phi \text{ iff } \forall w' (w D_{agt:\alpha} w' \Rightarrow M, w' \models \phi) \\
& M, w \models E_{agt:\alpha}^+ \phi \text{ iff} \\
(1) & \forall w' (w D_{agt:\alpha} w' \Rightarrow M, w' \models \phi) \text{ and} \\
(2) & \exists w'' (w D_{agt:\neg\alpha} w'' \text{ and } M, w'' \models \neg\phi)
\end{aligned}$$

The condition (1) expresses that it is sufficient in w that the agent agt does α to obtain the effect ϕ (it could be expressed by: $M, w \models \text{Does}_{agt:\alpha} \phi$), and the condition (2) expresses that it was necessary in w that the agent agt did α to obtain the effect ϕ .

The semantics of the action operator *Done* (respectively E) is defined in function of the semantics of the action operator *Does* (respectively E^+)⁵.

The intuition of these definitions is that if in a world w we have, for example, $E_{i:\alpha} \phi$, then there must exist a previous world w_1 where the action α has started and where we have $E_{i:\alpha}^+ \phi$. However, the situation may be more complex because the proposition ϕ may also contain an action operator, for example we may have $\phi = E_{j:\beta} \psi$. Then, the meaning of $E_{i:\alpha} (E_{j:\beta} \psi)$ is that in w i has brought it about that j has brought it about that ψ . That means that there must exist a world w_1 where i is going to bring it about that in a further world w_2 j is going to bring it about that in w the proposition ψ holds. That is, in w_1 we have $E_{i:\alpha}^+ (E_{j:\beta}^+ \psi)$, which has the effect that in w_2 we have $E_{j:\beta}^+ \psi$, which itself has the effect that in w we have ψ .

Since there is no fixed limitation in nesting the action operators we have the following recursive definitions of the semantics of the operators E and *Done*.

To avoid to repeat similar definitions we adopt the following notations: A denotes an action operator which may be either *Done* or E , and A^+ denotes *Does* (respectively E^+) if A denotes *Done* (respectively E). Moreover, a is used to denote $agt : \alpha$. Then, we have:

$$M, w \models A_a \phi \text{ iff } \exists w_1 (\exists w_2 (w_1 D_a w_2 \text{ and } \text{Path}(\phi, w_2, w)) \text{ and } M, w_1 \models A_a^+ T(\phi))$$

⁴ In [25] the worlds u' related to the world u where we are are defined in that way: "we must consider all those hypothetical situations u' in which the agent does as much as he does in u ", and that is the reason why the accessibility relation D is assumed to be reflexive by Pörn.

⁵ The operators *Done* and E are added to the language because "counts as" statements refer to actions that have been performed.

The formulas denoted by $Path(\phi, w_2, w)$ and $T(\phi)$ are recursively defined in the same way for the operators: $Done_a$ and E_a , and these operators can be nested and mixed without any fixed limitation. We have:

- If ϕ is not of the form: $A_a\phi_n$, then
 $Path(\phi, w_n, w) \stackrel{\text{def}}{=} (w_n = w)$ and $T(\phi) \stackrel{\text{def}}{=} \phi$.
- If ϕ is of the form $A_a\phi_n$, then
 $Path(\phi, w_n, w) \stackrel{\text{def}}{=} \exists w_{n+1}(w_n D_a w_{n+1}$ and $Path(\phi_n, w_{n+1}, w))$
and $T(\phi) \stackrel{\text{def}}{=} A_a^+ T(\phi_n)$.

The intuition of the formulas $\exists w_1(\exists w_2(w_1 D_a w_2$ and $Path(\phi, w_2, w))$ is that the sequence of nested actions which have been performed has started in the world w_1 , and if there is no nested operator the world w_2 is the world w where we are after action a performance.

This notion of path by itself is not new in Dynamic Logic and has been used by several authors like Segerberg in [31, 33] and Dignum et al. in [22]. However, none of them have defined how to reconstruct this path backward as it is done here with the recursive definition of paths.

It can be easily proved that we have the following properties for the action operators.

- (DD) $\models Does_a(Done_a)$
- (ED) $\models E_a^+ \phi \rightarrow Does_a \phi$
- ($\neg N+$) $\not\models E_a^+ \top$

The intuition of (ED) is that if action a causes ϕ , then after a performance ϕ holds.

- (DO) $\models Done_a \phi \rightarrow \phi$
- (E) $\models E_a \phi \rightarrow \phi$
- ($\neg N$) $\not\models E_a \top$

The intuition of (DO) and (E) is that after a performance ϕ holds. The difference between these operators is that $Done$ does not express causality.

- (EE \wedge) $\models E_a^+ \phi \wedge E_a^+ \psi \rightarrow E_a^+(\phi \wedge \psi)$

Property (EE \wedge) is quite intuitive.

- (EDE) $\models E_a^+ \phi \wedge Does_a \psi \rightarrow E_a^+(\phi \wedge \psi)$

The property (EDE) may seem to be counter intuitive, because we are inclined to think that if a causes $\phi \wedge \psi$, then a causes ϕ and a causes ψ , which is not the case in general as shows the following property.

- ($\neg EE$) $\not\models E_a^+(\phi \wedge \psi) \rightarrow E_a^+ \psi$

See, for example, the case where ψ is a tautology. In that case, if we have $E_a^+ \phi$, we also have $E_a^+(\phi \wedge \psi)$ and, of course, we don't have $E_a^+ \psi$.

- (EE \vee) $\models E_a^+(\phi \wedge \psi) \rightarrow E_a^+ \phi \vee E_a^+ \psi$

The intuition of property (EE \vee) is that if a causes $\phi \wedge \psi$, then a causes either ϕ or ψ . If we would not have this property it might happen that action a causes $\phi \wedge \psi$ and a guarantees neither the truth of ϕ nor the truth of ψ .

3.4 Role

In [25], Pörn proposes this definition of a role: “Because of their prevalence in normative systems clusters of norms organized in this way deserve a name on their own. We shall call them role structure because in terms of them it is possible to define the sociological notion of a role.” . In other words, if, in a normative system, we repeatedly need to talk about the set of individuals to whom a given set of norms applies, it is convenient to select a name for this set of norms, and the set of norms which is referred to by this name is called a “role”.

However, this definition raises a difficulty if the set of norms contains norms defining some kinds of institutional powers. In particular, if an institutional power says that the fact that the role holder has performed some action acting as the role holder counts as something else, we see that we refer to the role in a norm which contributes to the definition of the role itself. That is, we have the role name which occurs both in the *definiens* and in the *definendum* and that means that we have a circular definition.

To avoid this circularity problem we shall say that a role refers to a set of norms, but we do not say that a role is defined by a set of norms. In this approach there is no more the circularity problem and the properties of a role can be defined with the predicate *Holds* which is defined below.

Holds(h, r, i, s): human agent h holds the role r in the organization of the institutional agent i in the context of the institution s ⁶.

For example, a situation where for any given h who holds a given role r in a given institutional agent i in the context of a given institution s , the fact that h has performed a given action α counts as the fact that i has performed a given action β can be represented in a semi-formal way by the following formula.

$$\forall h(\text{Holds}(h, r, i, s) \rightarrow (\text{Done}_{h:r:\alpha}) \Rightarrow_s \text{Done}_{i:\beta})$$

We insist on the fact that this formula is not an axiom schema, it is an example of norm of the kind counts as, and the symbols r, i, s, α and β are constant symbols. The variable symbol h is universally quantified because this kind of norm is not defined for a specific given individual.⁷

The fact that acting of some role holder requires to hold this role and to fulfill some given conditions during the performance of the action can be represented by the following formula.

$$\forall h(\text{Done}_{h:r:\alpha} \rightarrow (\text{Holds}(h, r, i, s) \wedge \text{act.cond}_{\alpha,r,i,s}))$$

where $\text{act.cond}_{\alpha,r,i,s}$ denotes the specific conditions which have to be fulfilled to recognize that the agent has done the action α acting as holder of the role r . Here again r, i, s and α denote constant symbols.

⁶ A more formal definition of the predicate *Holds* has been presented by Demolombe and Louis in [9].

⁷ This formula is “semi-formal” because it contains a universally quantified variable h which occurs both as an argument of the predicate *Holds* and as an index of the modal operator *Done*. To give a formal semantics to this kind of formula is out of the scope of this paper because it would require to go into too long technical details.

4 Application of the logical framework

In this section the logical framework defined in section 3 is applied to the case study presented in section 2.

The counts as statements are formally represented as follows.

- (CT1) $Obg_T Done_{I_1:\alpha} \Rightarrow_S Obg_S Done_{H_1:r_1:\beta}$
- (CT2) $Done_{H_1:r_1:\omega} \Rightarrow_S Obg_S Done_{H_2:r_2:\gamma}$
- (CT3) $Done_{H_1:r_1:\beta} \Rightarrow_S Done_{I_1:\alpha}$
- (CT4) $Done_{H_1:r_1:\beta} \Rightarrow_T Done_{I_1:\alpha}$
- (CT5) $Done_{H_2:r_2:\gamma} \Rightarrow_S Done_{H_1:r_1:\beta}$
- (CT6) $Done_{H_2:r_2:\gamma} \Rightarrow_T Done_{H_1:r_1:\beta}$
- (CT7) $E_{H_2:r_2:\gamma_1} E_{S_1:\gamma_2} Done_{S_2:\gamma_3} \Rightarrow_S Done_{H_2:r_2:\gamma}$
- (CT8) $E_{H_2:r_2:\gamma_1} E_{S_1:\gamma_2} Done_{S_2:\gamma_3} \Rightarrow_T Done_{H_2:r_2:\gamma}$

The counts as norms (CT1) to (CT8) are parts of the definitions of the roles r_1 and r_2 . As a matter of simplification we have considered instances of these norms for the given human agents H_1 and H_2 while in general these norms are defined for any agent who holds these roles. For instance, instead of (CT2) we should have :

(CT2') $\forall h \forall h' (Holds(h, r_1, I_1, S) \wedge Holds(h', r_2, I_1, S) \rightarrow (Done_{h:r_1:\omega} \Rightarrow_S Obg_S Done_{h':r_2:\gamma}))$

and from $Holds(H_1, r_1, I_1, S)$ and $Holds(H_2, r_2, I_1, S)$ we could infer (CT2).

For the same reasons, instead of (CT4) we should have:

(CT4') $(\exists h Done_{h:r_1:\beta}) \Rightarrow_T Done_{I_1:\alpha}$

because the institution T does not need to know who has performed β , provided β has been performed by someone who is acting as a r_1 role holder.

Now, it is assumed that we have :

(A1) $Obg_T Done_{I_1:\alpha}$

Then, we have :

(1) $D_S Obg_S Done_{H_1:r_1:\beta}$ from (A1), (CT1), (D) and (C)

(2) $Obg_S Done_{H_1:r_1:\beta}$ from (1) and (R)

The fact that H_1 has decided to order to H_2 to do γ is represented by :

(A2) $Done_{H_1:r_1:\omega}$

Then, we have :

(3) $D_S Obg_S Done_{H_2:r_2:\gamma}$ from (A2), (CT2), (D) and (C)

(4) $Obg_S Done_{H_2:r_2:\gamma}$ from (3) and (R)

The fact that H_2 has decided to run S_1 and S_2 is formally represented by :

(A3) $E_{H_2:r_2:\gamma_1} E_{S_1:\gamma_2} Done_{S_2:\gamma_3}$

Then, we have :

(5) $D_T Done_{H_2:r_2:\gamma}$ from (A3), (CT8), (D) and (C)

(6) $D_T Done_{H_1:r_1:\beta}$ from (5), (CT6) and (D)

(7) $D_T Done_{I_1:\alpha}$ from (6), (CT4) and (D)

The consequence (7) shows that the obligation (A1) $Obg_T Done_{I_1:\alpha}$ is fulfilled, and it is fulfilled thanks to the fulfillment of the obligations (2) and (4). Indeed, we have :

(8) $D_S Done_{H_2:r_2:\gamma}$ from (A3), (CT7), (D) and (C)

(9) $D_S Done_{H_1:r_1:\beta}$ from (8), (CT5) and (D)

The sentences (A1), (2) and (4) show how obligations are propagated from I_1 to H_1 and to H_2 , and the sentences (A3), (5), (6) and (7) show how the actions done by H_2 , S_1 and S_2 count as actions done by H_1 and I_1 , and justify the fulfillment of obligation (A1). Notice that it would be wrong to accept instead of (CT7) the counts as norm :

$$E_{S_1:\gamma_2} Done_{S_2:\gamma_3} \Rightarrow_S Done_{H_2:r_2:\gamma}$$

because the software agents S_1 and S_2 might be run by another human agent than H_2 . Of course, the same comment holds for (CT8). Roughly speaking we have to make explicit who is the human agent who has triggered the software agents.

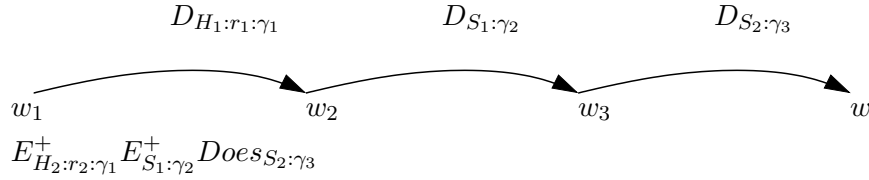


Fig. 2. Interpretation of action operators.

We take opportunity of this case study to show how the satisfiability conditions for the action operator E are evaluated for the action: $E_{H_2:r_2:\gamma_1} E_{S_1:\gamma_2} Done_{S_2:\gamma_3}$. The general form of this formula is: $A_a \phi_1$, where $A = E$ and $a = H_2 : r_2 : \gamma_1$, and where we have:

$$\begin{aligned} \phi_1 &= E_{S_1:\gamma_2} Done_{S_2:\gamma_3} \\ \phi_2 &= Done_{S_2:\gamma_3} \\ \phi_3 &= \top \end{aligned}$$

Then, the general form of the satisfiability condition is instantiated by:

$$M, w \models E_{H_2:r_2:\gamma_1} \phi_1 \text{ iff } \exists w_1 (\exists w_2 (w_1 D_{H_2:r_2:\gamma_1} w_2 \text{ and } Path(\phi_1, w_2, w)) \text{ and } M, w_1 \models E_{H_2:r_2:\gamma_1}^+ T(\phi_1))$$

From T definition we have:

$$\begin{aligned} T(\phi_1) &= E_{S_1:\gamma_2}^+ T(\phi_2) \\ T(\phi_2) &= Does_{S_2:\gamma_3} T(\phi_3) \\ T(\phi_3) &= \top \end{aligned}$$

From $Path$ definition we have:

$$\begin{aligned} Path(\phi_1, w_2, w) &= \exists w_3 (w_2 D_{S_1:\gamma_2} w_3 \text{ and } Path(\phi_2, w_3, w)) \\ Path(\phi_2, w_3, w) &= \exists w_4 (w_3 D_{S_2:\gamma_3} w_4 \text{ and } Path(\phi_3, w_4, w)) \\ Path(\phi_3, w_4, w) &= (w_4 = w) \end{aligned}$$

Then, the final rewriting of these conditions is:

$$\begin{aligned} M, w \models E_{H_2:r_2:\gamma_1} E_{S_1:\gamma_2} Done_{S_2:\gamma_3} \text{ iff} \\ \exists w_1 (\exists w_2 (w_1 D_{H_2:r_2:\gamma_1} w_2 \text{ and } \exists w_3 (w_2 D_{S_1:\gamma_2} w_3 \text{ and } w_3 D_{S_2:\gamma_3} w)) \text{ and } \\ M, w_1 \models E_{H_2:r_2:\gamma_1}^+ E_{S_1:\gamma_2}^+ Does_{S_2:\gamma_3}) \end{aligned}$$

The sequence of worlds: w_1, w_2, w_3 and w can be seen as the past history that has lead to obtain $E_{H_2:r_2:\gamma_1} E_{S_1:\gamma_2} Done_{S_2:\gamma_3}$ in w (see figure 2).

Indeed, we have: $E_{H_2:r_2:\gamma_1}^+ E_{S_1:\gamma_2}^+ Does_{S_2:\gamma_3}$ in w_1 , $E_{H_2:r_2:\gamma_1} E_{S_1:\gamma_2}^+ Does_{S_2:\gamma_3}$ in w_2 , $E_{H_2:r_2:\gamma_1} E_{S_1:\gamma_2} Does_{S_2:\gamma_3}$ in w_3 and $E_{H_2:r_2:\gamma_1} E_{S_1:\gamma_2} Done_{S_2:\gamma_3}$ in w .

5 Comparison with other works

There is a limited number of works about the logical formalization of the counts as operator ⁸ even if some authors have proposed formalizations which are not based on formal logic (see [13]).

In [14] Gelati et al. have proposed a variant of Jones and Sergot formalization which is based on a defeasible conditional operator, denoted by \Rightarrow , and which can be used for any defeasible normative connection. Then, the counts as connection takes the form: $(\phi \Rightarrow D_S \psi) \wedge (D_S \phi \Rightarrow D_S \psi)$.

Another significant variant has been proposed by Grossi in [16, 17] (a similar approach has been proposed by Lorini et al. in [23]). Grossi has presented several definitions of this operator and all of them share the formal property that the antecedent and the consequent of the counts as operators appear as the antecedent and the consequent of a material implication which is in the scope of modal operators. For example, the "proper classificatory" operator $\phi \Rightarrow_S^{cl+} \psi$ ⁹ is defined as: $[s](\phi \rightarrow \psi) \wedge \neg[u](\phi \rightarrow \psi)$, where the intuition of $[u](\phi \rightarrow \psi)$ is that $\phi \rightarrow \psi$ holds in the context of any institution.

The first unexpected consequence is that, since the material implication $\phi \rightarrow \psi$ is logically equivalent to $\neg\phi \vee \psi$, and disjunction is commutative, the antecedent and the consequent have the same "status". That is, there is no distinction between brute facts and institutional facts, and we can find many examples where $\phi \Rightarrow_S^{cl+} \psi$ entails $\neg\psi \Rightarrow_S^{cl+} \neg\phi$.

Another weakness of this formalization is that from a brute fact ϕ we cannot infer any institutional consequence because brute facts are not in the scope of the modal operator $[s]$. For instance, in our case study, from the brut fact: $Done_{H_1:r_1:\omega}$ and the counts as assumption: $Done_{H_1:r_1:\omega} \Rightarrow_S Obg_S Done_{H_2:r_2:\gamma}$ we cannot infer with these definitions the institutional fact represented by: $D_S Obg_S Done_{H_2:r_2:\gamma}$.

The logical formalization of causality or agency has deserved a large number of works. A survey of these works can be found in [32] (see also [30]) and [1]. The formalization proposed by von Wright [34] has been a reference for many authors. What has been presented in this paper takes inspiration both from von Wright and from further works by Kanger [21] and Pörn [25]. In [8] Demolombe and Jones have analyzed the relationships between the bringing it about action operators and deontic operators. Hilpinen in [18] has proposed a more refined characterization of the counterfactual conditions.

⁸ See also in [28] the distinction between the concepts of "counting as" and "emergence" proposed by Sartor.

⁹ We have slightly changed the notation to make easier the comparison with Jones and Sergot's definition.

In [19] Horty and Belnap have defined the so called STIT operators. Their basic idea is that at a given moment an agent can chose among a given set of action options and the option he has chosen causes ϕ iff for all the histories that conform the same choice ϕ is obtained and there exists another choice such that for some history which conforms this latter choice ϕ is not obtained. One of the reasons why we have not adopted this approach is that this formalization is not appropriate to represent situations where agents are software agents whose actions are not the consequence of a deliberative choice in its genuine sense.

In the deliberative STIT approach it is inconsistent¹⁰ to say that an agent i has made a choice whose effect is that an other agent j has made some particular choice. At the opposite, with the bringing it about operator we have defined it is not inconsistent to say, for example, that the human agent H_2 by doing γ_1 has brought it about that the software agent S_1 by doing γ_2 has brought it about that the software agent S_2 has done γ_3 .

For the formalization of the obligations we have adopted the Standard Deontic Logic which is the simplest deontic logic. One of the features which is ignored in this logic is the temporal dimension which is quite relevant in the case of obligations about actions. Indeed, an obligation to do an action is not completely defined if we ignore the deadline which is imposed to do the action.

There are few works about the logical formalization of roles. Cuppens in [5] has proposed such a formalization where roles are seen as "virtual agents" and, for example, what is permitted to do is what this virtual agent does in some ideal world. In [3, 26, 24] Carmo, Pacheco and Santos have defined a role as a set of obligations, permissions and prohibitions as proposed by Pörn in [25]. In [9] Demolombe and Louis have extended this definition to institutional powers that allow us to create obligations and permissions. Santos and Pacheco in [24] have also defined action operators of the kind to bring it about which are indexed by pairs of agents and roles as we did in this paper though their definition of these operators are closer to Pörn's definition. They also require the constraint that if an agent is acting as holder of a role, then he holds this role. However, as mentioned in section 3.3, we also have to impose additional constraints because even if an agent holds a role he is not acting as holder of this role for every actions he does. These constraints are called "enacting" constraints by Dastani, Dignum and Dignum in [6].

In [10] Dignum and Dignum have defined a logical framework to model organizations. This framework extends the temporal logic CTL with modal operators to represent the concepts of agents' ability and activity. Organizations are defined in terms of their state, desire and ability, plus responsibility dependencies. Though this framework has a large expressive power it does not consider the three kinds of agents which have been analyzed here. In [11] they have proposed three levels that are called: "abstract", "concrete" and "implementation", but these levels have completely different meaning.

¹⁰ In [2] it has been formally proved that this inconsistency follows from the independence constraint about agents' choices.

In [7] de Lima et al. combine in the same logical framework the STIT operator and obligation operator in order to formalize the notion of responsibility. However, the notion of institutional agent and the counts as operator are ignored.

6 Conclusion

The case study analysis has shown how obligations about institutional agents can be propagated to human and software agents, and how norms of the kind "counts as" justify that actions performed by software agents and by human agents fulfill these obligations in the context of a given institution.

It has also been shown that this propagation is realized in different ways depending on the kind of agents. From institutional agents to human agents it is done through counts as statements defining relations between obligations about institutional agents and obligations about human agents. From human agents to human agents it is done by orders that creates obligations. From human agents to software agents it is done by human agent actions that cause software agent actions.

The foundation of the counts as norms which justify the creation of obligations differ depending on the kind of agents.

In the case of norms defining relations between software agent actions and human actions the foundation is the causal relationship between these actions. In the case of norms defining relations between human agent actions and other human agent actions or institutional agent actions it is based on the fact that these agents are acting as holder of some roles.

Since the most relevant concepts which are involved in the description of this kind of case study are: counts as, causality, obligation and role, we have proposed a logical framework only for these concepts. For the counts as we have adopted the Jones and Sergot's logical framework. For the causality we have proposed a new logic which takes inspiration from Pörn and von Wright logics. The formalization of obligations is an extension of Standard Deontic Logic with some additional properties that link obligation and the Jones and Sergot's D_S operator. For the formalization of roles we did not require a detailed analysis.

Some of the modal operators are normal modal operators and their semantics is defined by accessibility relations, while the counts as operator is a classical modal operator whose semantics is defined by functions. It is not problematic to integrate all these operators in a uniform semantics since, as it is shown in [4], normal operators can be defined as specific cases of classical operators.

The presented logical framework is defined in the semantics because we need clear definitions of the concepts before defining an axiomatics that allows us to derive consequences from a set of assumptions. The definition of such axiomatics should deserve further works.

In the paper we have focused on fulfillment of obligations to do actions. A similar analysis remains to be done for the violations which raise non trivial issues in the case of institutional agents. Indeed, an obligation about an institutional agent of the form $Obg_S Done_{I_1:\alpha}$ is violated iff we have $D_S \neg Done_{I_1:\alpha}$. How can

we guarantee that in the context of institution S it is not the case that I_1 has done α ? That cannot be based on direct observations since I_1 is not a concrete agent and $Done_{I_1:\alpha}$ is an institutional fact. Can we infer $D_S \neg Done_{I_1:\alpha}$ from other brute facts or institutional facts and counts as norms? That is not clear because a counts as norm of the form $\phi \Rightarrow_S \neg Done_{I_1:\alpha}$ seems to be rather odd. Can we infer $D_S \neg Done_{I_1:\alpha}$ from the fact that for none of the norms of the form $\phi \Rightarrow_S Done_{I_1:\alpha}$ the antecedent ϕ holds? That would implicitly require an inference rule of the kind "negation as failure" whose semantics raises other kinds of problems.

Finally the same analysis remains to be done for prohibition fulfillment. Indeed, if prohibitions are defined as usual, the prohibition for I_1 to have done α is represented by $Obj_S(\neg Done_{I_1:\alpha})$ and it is fulfilled iff we have $D_S \neg Done_{I_1:\alpha}$.

References

1. L. Aqvist. Old foundations for the logic of agency and action. *Studia Logica*, 72, 2002.
2. P. Balbiani, A. Herzig, and N. Troquard. Alternative axiomatics and complexity of deliberative stit theories. *Journal of Philosophical Logic*, 37(4):387–406, 2008.
3. J. Carmo and O. Pacheco. Deontic and action logics for organized collective agency, modeled through institutionalized agents and roles. *Fundamenta Informaticae*, 48:129–163, 2001.
4. B. F. Chellas. *Modal Logic: An introduction*. Cambridge University Press, 1988.
5. F. Cuppens. Roles and Deontic Logic. In A. J. I. Jones and M. Sergot, editors, *Second International Workshop on Deontic Logic in Computer Science*, Oslo, Norway, 1994.
6. M. Dastani, V. Dignum, and F. Dignum. Role-assignment in open agent societies. In *Proceedings of the Second International Conference on Autonomous Agents and Multiagent Systems*, 2003.
7. T. de Lima, L.M.M. Royakkers, and F. Dignum. Behaving responsible in multi-agent worlds. In K. Decker et al., editor, *8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, 2009.
8. R. Demolombe and A.J. Jones. Actions and normative positions. A modal-logical approach. In D. Jacquette, editor, *Companion to Philosophical Logic*. Blackwell, 2002.
9. R. Demolombe and V. Louis. Norms, institutional power and roles: toward a logical framework. In F. Esposito, Z. W. Ras, D. Malerba, and G. Semeraro, editors, *Foundations of Intelligent Systems*. Springer, LNAI 4203, 2006.
10. V. Dignum and F. Dignum. A logic for agent organizations. In *Multi-Agent Systems: Semantics and Dynamics of Organizational Models*, pages 220–241. IGI Publications, 2009.
11. V. Dignum, J. Vazquez-Salceda, and F. Dignum. OMNI: Introducing Social Structure, Norms and Ontologies into Agent Organizations. In R. H. Bordini, M. Dastani, and J. Dix, editors, *Proceedings of the International Workshop Programming Multi-Agent Systems (ProMAS 2004) LNAI 3346*. Springer, 2005.
12. M. Esteva, J. Rodríguez-Aguilar, J. Lluís Arcos, C. Sierra, P. Noriega, and B. Rosell. Electronic Institutions Development Environment. In *Proceedings of the 7th international joint conference on Autonomous Agents and Multiagent Systems (AAMAS08)*. 2008.

13. N. Fornara, F. Viganò, and M. Colombetti. Agent communication and institutional reality. In R. van Eijk, M. Huget, and F. Dignum, editors, *Developments in Agent Communication*. Springer Verlag LNAI 3396, 2005.
14. J. Gelati, G. Governatori, A. Rotolo, and G. Sartor. Declarative power, representation, and mandate: A formal analysis. In T. Bench-Capon, A. Daskalopulu, and R. Winkels, editors, *Frontiers in Artificial Intelligence and Applications, Number 89*. IOS Press, 2002.
15. G. Governatori, M. Dumas, A. H.M. ter Hofstede, and P. Oaks. A formal approach to protocols and strategies for (legal) negotiation. In H. Prakken, editor, *Proceedings of the 8th International Conference on Artificial Intelligence and Law*. ACM Press, 2001.
16. D. Grossi. *Designing Invisible Handcuffs. Formal Investigations in Institutions and Organizations for Multi-Agent Systems*. PhD thesis, Utrecht University, 2007.
17. D. Grossi, J-J. Ch. Meyer, and F. Dignum. The Many Faces of Counts-as: A Formal Analysis of Constitutive Rules. *Journal of Applied Logic*, 6, 2008.
18. R. Hilpinen. On Action and Agency. In E. Ejerhed and S. Lindstrom, editors, *Logic, Action and Cognition: Essays in Philosophical Logic*. Kluwer, 1997.
19. J.F. Horty and N. Belnap. The deliberative STIT: a study of action, omission, ability, and obligation. *Journal of Philosophical Logic*, 24:583–644, 1995.
20. A. J. Jones and M. Sergot. A formal characterisation of institutionalised power. *Journal of the Interest Group in Pure and Applied Logics*, 4(3), 1996.
21. S. Kanger. New foundations of ethical theory. In R. Hilpinen, editor, *Deontic logic*, pages 36–58. D. Reidel Publishing Company, 1983.
22. M. Kuiper, J-J.Ch. Meyer, and F.P.M. Dignum. An investigation into deontics of durative actions. In P. McNamara and H. Prakken, editors, *Proceedings of Deontic Logic in Computer Science (DEON '98)*, 1998.
23. E. Lorini, D. Longin, B. Gaudou, and A. Herzig. The Logic of acceptance: grounding institutions on agents attitudes. *Journal of Logic and Computation*, 19(6), 2009.
24. O. Pacheco and F. Santos. Delegation in a role-based organization. In Alessio Lomuscio and Donald Nute, editors, *Deontic Logic in Computer Science*, LNCS 3065. Springer, 2004.
25. I. Pörn. Action Theory and Social Science. Some Formal Models. *Synthese Library*, 120, 1977.
26. F. Santos and O. Pacheco. Specifying and reasoning with institutional agents. In *Proceedings of ICAIL*, 2003.
27. M. Sardis and G. A. Vouros. Electronic institutions infrastructure for e-chartering. In A. Artikis, G. M. P. O'Hare, K. Stathis, and G. A. Vouros, editors, *ESAW*, LNCS 4995. Springer, 2008.
28. G. Sartor. *Legal Reasoning: A Cognitive Approach to the Law*. Springer, Berlin, 2005.
29. J. R. Searle. *Speech Acts: An essay in the philosophy of language*. Cambridge University Press, New-York, 1969.
30. K. Segerberg. Bringing it about. *Journal of Philosophical Logic*, 18:327–347, 1989.
31. K. Segerberg. Getting started: beginnings in the logic of action. *Studia Logica*, 51:347–378, 1992.
32. K. Segerberg. Outline of a logic of action. In F. Wolter, H. Wansing, W. de Rijke, and M. Zakharyashev, editors, *Advances in Modal Logic, Volume 3*. World Scientific Publishing Co., 2002.

33. K. Segerberg. Some Meinong/Chisholm thesis. In K. Segerberg and K. Sliwinski, editors, *Logic, Law, Morality. A festrift in honor of Lennart Aqvist*, volume 51, pages 67–77. Uppsala Philosophical Studies, 2003.
34. G. H. von Wright. *Norm and Action*. Routledge and Kegan, 1963.
35. F. Lopez y Lopez, M. Luck, and M. d'Inverno. Normative agent reasoning in dynamic societies. In *Proceedings of the Third International Conference on Autonomous Agents and Multi-Agent Systems*. IEEE Computer Society, 2004.