

Collective obligations, commitments and individual obligations: a preliminary study

Laurence Cholvy and Christophe Garion
ONERA Toulouse
2 avenue Edouard Belin
BP 4025, 31055 Toulouse Cedex 4
email: {cholvy, garion}@cert.fr

Abstract

A collective obligation is an obligation directed to a group of agents so that the group, as a whole, is obliged to achieve a given task.

The problem investigated here is to study the impact of collective obligations to individual obligations, i.e. obligations directed to single agents of the group. The groups we consider do not have any particular hierarchical structure nor have an institutionalized representative agent. In this case, we claim that the derivation of individual obligations from collective obligations depends on several parameters among which the ability of the agents (i.e. what they can do) and their own personal commitments (i.e. what they are determined to do). As for checking if these obligations are fulfilled or not, we need to know what are the actual actions performed by the agents.

This present paper addresses these questions in the rather general case when the collective obligations are conditional ones.

1 Introduction

This paper studies the relation between collective obligations directed to a group of agents and the individual obligations directed to the single agents of the group. We study this relation in the case when the group of agents is not structured by any hierarchical structure and has no representative agent like in [4].

According to Royakkers and Dignum, [8], a collective obligation is an obligation directed to a group of individuals i.e. a group of agents. For instance (and this is an example given by Royakkers and Dignum) when a mother says: *Boys, you have to set the table*, she defines an obligation aimed at the group of her boys.

A collective obligation addressed to a group of agents is such that this group, as a whole, is obliged to achieve a given task. This comes to say that a given task is assigned as a goal to the group as a whole. In the mother's example, the goal assigned to the boys is to set the table and the mother expects that the table will be set by some actions performed by her boys. Whether only one of her boys or all of them will bring it about that the table is set is not specified by the mother.

In particular, one must notice that in the example, the mother does not oblige each of her boys to set the table. This shows the difference between collective obligations and what Royakkers and Dignum call "restricted general obligations" which are addressed to every member of the group. For instance, *Boys, you have to eat properly* is not a collective obligation but a restricted general obligation directed to every mother's boy.

What is particularly interesting with collective obligations is to understand their impact on the individual obligations of the agents in the group i.e., to understand when and how the collective obligations are translated into individual obligations. In the mother's example, will the eldest boy have to carry the forks and knives, the second the glasses and the youngest the plates ? or will the youngest have to carry everything ?

One can notice that when the mother directs the collective obligation to her boys, she does not direct (even implicitly) individual obligations to some or all of her boys. More generally, we think that when an agent directs a collective obligation to a group, it does not define individual obligations to some or all agents of the group. The consequence is that, in case of violation of the collective obligation, the only possible responsible towards the one who directed the obligation, is the group as a whole: no precise agent can

be responsible of the violation of a collective obligation in front of the agent who directed that collective obligation.

However, we think that when the agents of a group with no hierarchical structure, receive a collective obligation, they may coordinate themselves to provide a plan (or a task allocation), by committing themselves to make some actions. These commitments imply individual obligations that some agents must satisfy.

Understanding how the collective obligations are translated into individual obligations is the problem which is investigated here. We claim that the derivation of individual obligations from collective obligations depends on several parameters among which the ability of the agents (i.e. what each agent can do) and their own personal commitments (i.e. what each agent is determined to do). Latter on, by examining the actual actions of each agent of the group, one can check if these obligations are satisfied or violated.

For instance, if all the boys keep on watching TV (thus, do not set the table) then the collective obligation is violated. Or, even if the two youngest boys carry the forks, the knives and the plates, if the eldest one, who is the tallest and the only one who can take the glasses, does not take the glasses, the collective obligation will be violated too. As said previously, the whole group is responsible of the violation of the collective obligation. This can be questionable, particularly by the two youngest boys in the last case since they will be all punished because of the eldest's actions. However we will show that, in this case, the eldest can be taken as responsible by the group because he was the only one able to take the glasses.

This present paper addresses the question of the translation of a collective obligation into individual obligations in the rather general case when the collective obligations are conditional ones. Roughly speaking, a conditional obligation is an obligation which applies when a given condition is true. For instance, *If it is sunny, set the table in the garden; else set it in the dinner-room* defines two conditional obligations: if it is sunny, the boys have to set the table in the garden but else, they have to set it in the dinner-room. In this work, we have chosen Boutilier's conditional preferences logic [3], [2], for representing conditional obligations.

So, this work assumes that a set of conditional obligations is directed to a group of agents. It also assumes a model of agents, describing each agent of that group by its knowledge about the current situation, its abilities and its commitments. It first defines a characterization of the obligations that the whole group have to satisfy. Then, grouping the agents according

to their ability, it then defines the obligations that such sub-groups have to satisfy. Finally, given the commitments of the agents, it defines their individual obligations. As for checking if these obligations are satisfied or not, we have to consider the results of the agents' actions.

This work is based on the work of Boutilier [3], who addresses some of these questions in the case of a single agent. In that paper, Boutilier assumes a set of conditional preferences expressing a goal for a single agent. He then describes a way to define the actual goals of the agent, given what it knows (or more exactly, what it believes) and given what it controls. Like Boutilier mentions it, this work can be applied to deal with obligations instead of goals. Our aim is to adapt Boutilier's work in the case of collective obligations. This will lead us to enrich the model of agents by considering their commitments.

This paper is organized as follows. Section 2 quickly presents Boutilier's work and in particular CO^* logic and the model of agent he considers. Section 3 adapts this work to the case of collective obligations and section 4 illustrates it on an example. Finally section 5 is devoted to a discussion.

2 A solution in the case of a single agent

This section quickly presents Boutilier's work in the case of a single agent. It first recall the semantics of the logic used by Boutilier then it recalls the model of agent he considers and its impact on the definition of goals.

2.1 CO and CO^* logics and conditional preferences

Given a propositional language $PROP$, Boutilier defines CO logic whose language extends $PROP$ with two primitive modal operators: \square and $\check{\square}$. Models of CO are of the form: $M = \langle W, \leq, \phi \rangle$ where W is a set of worlds, ϕ is a valuation function¹, and \leq is a total pre-order² on worlds, allowing one to express preference: $v \leq w$ means that v is at least as preferred as w .

Let $M = \langle W, \leq, \phi \rangle$ be a CO model. The valuation of a formula in M is given by the following definition:

Definition 1.

$M \models_w \alpha$ iff $w \in \phi(\alpha)$, for any propositional letter α .

$M \models_w \neg\alpha$ iff $M \not\models_w \alpha$ for any formula α .

¹i.e. $\phi : PROP \rightarrow 2^W$ such that $\phi(\neg\varphi) = W - \phi(\varphi)$ and $\phi(\varphi_1 \wedge \varphi_2) = \phi(\varphi_1) \cap \phi(\varphi_2)$

² \leq is reflexive, transitive and connected binary relation

$M \models_w (\alpha_1 \wedge \alpha_2)$ iff $M \models_w \alpha_1$ and $M \models_w \alpha_2$, if α_1 and α_2 are formulas.
 $M \models_w \Box \alpha$ iff for any world v such that $v \leq w$, $M \models_v \alpha$
 $M \models_w \check{\Box} \alpha$ iff for any world v such that $w < v$, $M \models_v \alpha$
 $M \models \alpha$ iff $\forall w \in W M \models_w \alpha$.

Thus, $\Box \alpha$ is true in the world w iff α is true in all the worlds which are at least as preferred as w . $\check{\Box} \alpha$ is true in the world w iff α is true in all the worlds which are less preferred than w . Dual operators are defined as usual:
 $\Diamond \alpha \equiv_{def} \neg \Box \neg \alpha$ and $\check{\Diamond} \alpha \equiv_{def} \neg \check{\Box} \neg \alpha$. Furthermore, Boutilier defines:
 $\check{\check{\Box}} \alpha \equiv_{def} \Box \alpha \wedge \check{\Box} \alpha$ and $\check{\check{\Diamond}} \alpha \equiv_{def} \Diamond \alpha \vee \check{\Diamond} \alpha$.

Let Σ be a set of formulas and α be a formula of CO . α is a logical consequence of (or deducible from) Σ iff any model which satisfies Σ also satisfies α . It is denoted as usual: $\Sigma \models \alpha$.

Boutilier then considers CO^* [2], a restriction of CO by considering a class of CO models in which any propositional valuation is associated with at least one possible world. The CO^* models are CO models M which satisfy: $M \models \check{\check{\Diamond}} A$, for any satisfiable formula A of $PROP$.

In the following, we only consider CO^* .

In order to express conditional preferences, Boutilier considers a conditional connective $I(-|-)$, defined by:

$$I(B|A) \equiv_{def} \check{\check{\Box}} \neg A \vee \check{\check{\Diamond}} (A \wedge \Box(A \rightarrow B))$$

$I(B|A)$ means that if A is true, then the agent ought to ensure that B .

An absolute preference is of the form: $I(A|\top)$ ³. It is denoted $I(A)$.

In order to determine its own goals, an agent must have a knowledge about the real world, or more exactly some beliefs about the real world. Boutilier thus introduces KB , a finite and consistent set of formulas of $PROP$, which expresses the beliefs the agent has about the real world. KB is called a knowledge base.

Given KB and given a model of CO^* , the most ideal situations are characterized by the most preferred worlds which satisfy KB . This is defined as follows:

Definition 2. Let Σ be a set of conditional preferences. Let KB be a knowledge-base. An ideal goal derived from Σ is a formula α of $PROP$ such that: $\Sigma \models I(\alpha|Cl(KB))$, where $Cl(KB) = \{\alpha \in PROP : KB \models \alpha\}$.⁴

³Where \top is any propositional tautology

⁴In fact, Boutilier uses a non monotonic logic to deduce the default knowledge of the agent. Here, in order to focus to ideal goals, we restrict to classical logic.

Example 1. Consider a propositional language whose letters are l (the door is lacquered) et s (the door is sanded down). Consider the two conditionals $I(l)$ and $I(\neg l|\neg s)$ which express that it is preferred that the door is lacquered but, if it is not sanded down, it is preferred that it is not lacquered. The possible worlds are: $w_1 = \{l, s\}$, $w_2 = \{\neg l, \neg s\}$, $w_3 = \{l, \neg s\}$, $w_4 = \{\neg l, s\}$ ⁵. Because of $I(l)$, the worlds w_1 and w_3 can be the most preferred ones. But, due to $I(\neg l|\neg s)$, w_3 cannot be one of the most preferred. Thus, w_1 is the only one most preferred, i.e. $w_1 \leq w_2$, $w_1 \leq w_3$ and $w_1 \leq w_4$. Furthermore, $w_3 \leq w_2$ is impossible because of $I(\neg l|\neg s)$. Thus $w_2 \leq w_3$. The models which satisfy $I(l)$ and $I(\neg l|\neg s)$ are thus the following:

$$\begin{aligned} M_1 : & w_1 \leq w_2 \leq w_3 \leq w_4 \\ M_2 : & w_1 \leq w_2 \leq w_4 \leq w_3 \\ M_3 : & w_1 \leq w_4 \leq w_2 \leq w_3 \end{aligned}$$

Assume first that $KB_1 = \{s\}$ (the door is sanded-down). Thus $Cl(KB_1) = \{s\}$. Ideal goals for the agent are α such that $\forall M M \models I(\alpha|s)$. l is thus an ideal goal for the agent: since the door is sanded-down, the agent has to lacquer it. Assume now that $KB_2 = \{\neg s\}$ (the door is not sanded-down). One can prove that $\neg l$ is now the ideal goal of the agent: since the door is not sanded-down, the agent must not lacquer it. This is questionable and discussed in the following.

2.2 Controllable, influenceable propositions and CK-goals

By definition 2, any formula α such that $M \models I(\alpha|Cl(KB))$ is a goal for the agent. Boutilier notes that this is questionable if KB is not “fixed” i.e. if the agent can change the truth value of some propositions in KB . For instance, in the second case of example 1, if the agent can sand-down the door, it would be preferable that he does so, and that he also lacquers it in order to achieve the most preferred situation.

Boutilier then suggests, in the definition of $Cl(KB)$, to take into account only the propositions whose truth value cannot be changed by some of the agent’s action.

⁵This way of denoting worlds is classical: for instance, $w_4 = \{\neg l, s\}$ is a notation to represent $w_4 \notin \phi(l)$ and $w_4 \in \phi(s)$

Furthermore, it may happen that some formulas α which are characterized by definition 2 define situations that the agent cannot achieve. Assume for instance that in the first case of example 1, the agent cannot lacquer the door (there is no more lacquer): lacquering cannot be a goal for the agent.

So, Boutilier introduces a partition of atoms of $PROP$: $PROP = C \cup \overline{C}$. C is the set of the atoms the agent controls (i.e. the atoms the agent can change the truth value) and \overline{C} is the set of atoms the agent does not control (i.e. the atoms the agent cannot change the truth value)

For instance, if the agent has got a sander and knows how it works, then we can consider that he controls the atom s . In any other case, we can consider that the agent does not control s .

Definition 3. For any set of propositional letters P , let $V(P)$ be the set of all the valuations of P . If $v \in V(P)$ and $w \in V(Q)$ with P and Q two disjoint sets, then $v;w \in V(P \cup Q)$ is the valuation extended to $P \cup Q$.

Definition 4. Let C and \overline{C} respectively be the set of atoms that the agent controls and the set of atoms that he does not control. A proposition α is *controllable* iff, for any $u \in V(\overline{C})$, there are $v \in V(C)$ and $w \in V(C)$ such that $v;u \models \alpha$ and $w;u \models \neg\alpha$. A proposition α is *influenceable* iff there are $u \in V(\overline{C})$, $v \in V(C)$ and $w \in V(C)$ such that $v;u \models \alpha$ and $w;u \models \neg\alpha$.

One can notice that for an atom, controllability and influenceability are equivalent notions. But this is not true for any non atomic propositions. Controllable proposition are influenceable, but the contrary is not true.

Definition 5. The set of the uninfluenceable knowledge of the agent is denoted $UI(KB)$ and is defined by:

$$UI(KB) = \{\alpha \in Cl(KB) : \alpha \text{ is not influenceable}\}$$

In a first step, Boutilier assumes that $UI(KB)$ is a complete set, i.e. the truth value of any element in $UI(KB)$ is known.⁶ Under this assumption, Boutilier then defines the notion of CK-goal:

Definition 6. Let Σ be a set of conditional preferences and KB a knowledge base such that $UI(KB)$ is complete. A proposition φ is a CK-goal (for the agent) iff $\Sigma \models I(\varphi|UI(KB))$ with φ controllable (by the agent).

Finally, Boutilier notices that goals can only be affected by atomic actions, so it is important to characterize the set of actions which are guaranteed to achieve each CK-goal. So he introduces the following notion:

⁶In a second step, Boutilier also examines the case when $UI(KB)$ is not complete. We will not focus on that case.

Definition 7. An *atomic goal set* is a set S of controllable atoms such that for any CK-goal φ , $\Sigma \models (UI(KB) \wedge S) \rightarrow \varphi$.

Example 1 (Continued). Consider again $\Sigma = \{I(l), I(\neg l | \neg s)\}$. Assume that $KB = \{\neg l, \neg s\}$ (the door is not sanded-down and not lacquered). Assume first that the agent can sand the door down and lacquer it. Then $UI(KB) = \emptyset$. Thus, $\{l, s\}$ is the atomic goal set of the agent: the agent has to sand the door down and to lacquer it. Assume now that the agent can only sand the door down but cannot lacquer it. Here, l is not controllable, thus $UI(KB) = \{\neg l\}$ and the agent has no atomic goal.

3 Collective obligations

Let us now consider that the conditional preferences are modeling collective obligations which are allocated to a group of agents $\mathcal{A} = \{a_1, \dots, a_n\}$. The problem we are facing now is to understand in which case these collective obligations define individual obligations and how to check if they are violated or not.

Following Boutilier, we will assume that each agent is associated with the atoms it controls and the atoms it does not control. But we extend that agency model by assuming that each agent is also associated with the atoms it commits itself to make true and the atoms it commits itself not to make true. These notions will be formalized latter, but intuitively, let us say that an agent commits itself to make an atom true if it expresses that it intends to perform an action that will make that atom true. An agent commits itself not to make an atom true if it expresses that it will perform no action that makes this atom true.

Assumption. In the following, the problem of determining individual obligations is studied assuming that the agents of the group have the same complete beliefs about the current world.

3.1 Obligations of the group

Here, we extend the notion of CK-goals to the case when there are several agents. For doing so, we first extend the notions of controllability and influenceability to a group of agents.

Let a_i be an agent of \mathcal{A} . Let C_{a_i} be the set of atoms which are controllable by a_i (i.e. atoms which a_i can change the truth value) and \overline{C}_{a_i} be the

set of the atoms that are not controllable by a_i . The extension of notions of controllability and influenceability for a group of agent is given by the following definition.

Definition 7. Let $C = \bigcup_{a_i \in \mathcal{A}} C(a_i)$ and $\overline{C} = PROP \setminus C$. A proposition α is *controllable* by the group \mathcal{A} iff, for any $u \in V(\overline{C})$, there are $v \in V(C)$ and $w \in V(C)$ such that $v;u \models \alpha$ and $w;u \models \neg\alpha$. A proposition α is *influenceable* iff there are $u \in V(\overline{C})$, $v \in V(C)$ and $w \in V(C)$ such that $v;u \models \alpha$ and $w;u \models \neg\alpha$.

This definition is obviously an extension of definition 4 to the multi-agent case.

Example 2. Consider a group of agents $\{a_1, a_2\}$ such that p is controllable by a_1 and r is controllable by a_2 . We can show that the proposition $(p \vee q) \wedge (r \vee s)$ is not controllable by $\{a_1, a_2\}$. Indeed, if q and s are both true, whatever the actions of a_1 and a_2 are, the proposition will remain true. However, $p \wedge r$ and $p \vee r$ are both controllable by $\{a_1, a_2\}$.

Since we assume the the agents share a common belief about the current world, we can still consider a knowledge base KB as a set of propositional formulas of $PROP$. And, like in the previous section, we assume that KB is complete.

Like in [5], we can show that, given a Knowledge Base KB , some propositions are true and uninfluenceable in KB even if they are influenceable according to definition 7.

Example 3. Consider a group $\{a_1, a_2\}$ such that p is controllable by a_1 and a_2 and q is not controllable neither by a_1 nor by a_2 . According to definition 7, the proposition $p \vee \neg q$, even if not controllable, is influenceable by the group. Let us now consider $KB = \{p, \neg q\}$. $p \vee \neg q$ is true in KB and, whatever the agents will do, will remain true. We will say that $p \vee \neg q$ is *uninfluenceable* in KB .

This leads to the extension of definition 7 as follows:

Definition 7 (continued). Given a knowledge-base KB , a proposition α is influenceable in KB iff there are $u \in V(\overline{C})$ such that $u \models KB$, $v \in V(C)$ and $w \in V(C)$ such that $v;u \models \alpha$ and $w;u \models \neg\alpha$.

Notice that the previous example shows that a proposition, which is a logical consequence of KB may be influenceable but uninfluenceable in KB .

We can thus introduce the following set:

Definition 8. Let $UI(KB)$ be the set of logical consequences of KB which are not influenceable by the group \mathcal{A} or not influenceable in KB by the group \mathcal{A} .

Definition 9. The group A has the obligation of φ towards the agent who directed the collective obligation iff $\Sigma \models I(\varphi|UI(KB))$ with φ controllable by \mathcal{A} . It is denoted $O_A\phi$.

This definition is obviously an extension of definition 6 to the multi-agent case. And we can check that it is also an extension, to the multi-agent case, of the notion of ideal obligations given in [5].

Thus, these obligations characterize the most preferred situation that the group A can achieve, given what is fixed and given what the whole group can control. But we can go further, by directing these obligations to the sub-groups that can really fulfill them.

Definition 10. Let ϕ a proposition. Let A_ϕ be the union of the minimal subsets of A which controls ϕ ⁷. We say that **the sub-group A_ϕ has the obligation of ϕ , towards A** iff $\Sigma \models I(\phi|UI(KB))$. It is denoted $O_{A_\phi}^A\phi$.

Thus, these obligations characterize the most preferred situation that the group A_ϕ (a group which is the union of all the minimal sub-groups which control ϕ) can achieve, given what is fixed.

Example 4. In the mother's example, let us assume that the mother obliges her boys to put the glasses and the forks on the table. In this case, the conditional preference which models this collective obligation is: $I(\text{glasses} \wedge \text{forks})$.

Assume that the first boy, *Paul*, is able to put the glasses on the table while the two others *John* and *Phil* are able to put the forks. Then we have:

$$O_{\{Paul, John, Phil\}}(\text{glasses} \wedge \text{forks})$$

$$O_{\{Paul\}}^{\{Paul, John, Phil\}} \text{glasses}$$

and

$$O_{\{John, Phil\}}^{\{Paul, John, Phil\}} \text{forks}$$

In other words, the group $\{Paul, John, Phil\}$ has the obligation (towards the mother) to put the glasses and the forks on the table.

The singleton $\{Paul\}$ has the obligation, towards the group $\{Paul, John, Phil\}$, to put the glasses on the table. And the group $\{John, Phil\}$ has the obligation, towards the group $\{Paul, John, Phil\}$, to put the forks on the table.

⁷We could also choose to define A_ϕ as some of the minimal subsets of A which controls ϕ . However, the whole study of the consequence of this alternative has not yet been done

3.2 Agents commitments

Given an atom it controls, an agent may have three positions. The agent can express that it will perform an action making this atom true. We will say that the agent commits itself to make that atom true. The agent can also express that it will perform no action making this atom true. We will say that it commits itself not to make that atom true. Finally, it can happen that the agent does not express that it will perform an action making the atom true nor expresses that it will perform no action making it true. In this case, the agent does not commit itself to make the atom true, and does not commit itself not to make it true.

These three positions are modeled by three subsets of the sets of atoms that an agent controls. $Com_{+,a_i} \subseteq C_{a_i}$ is the set of atoms a_i controls such that a_i commits itself to make them true. $Com_{-,a_i} \subseteq C_{a_i}$ is the set of atoms a_i controls such that a_i commits itself not to make them true. $P_{a_i} = C_{a_i} \setminus (Com_{+,a_i} \cup Com_{-,a_i})$ is the set of atoms a_i controls such that a_i does not commit to make them true nor commits not to make them true.

These sets are supposed to be restricted by the following constraints:

Constraint 1. $\forall a_i \in \mathcal{A}$ Com_{+,a_i} is consistent.

Constraint 2. $\forall a_i \in \mathcal{A}$ $Com_{+,a_i} \cap Com_{-,a_i} = \emptyset$

These two constraints are expressing a kind of consistency in the agent's model. By constraint 1, we assume that an agent does not commit itself to make something true and to make it false. By constraint 2, we assume that an agent does not commit itself to make an atom true and not to make it true.

Remark. The previous notions have been modeled in modal logic in [6], with two families of modal operators: C_i and E_i , $i \in \{1..n\}$. The operator E_i is the *stit* operator ([1], [7]). $E_i\phi$ intends to express that the agent a_i is seeing to it that ϕ . It is defined by the following axiomatics:

$$\begin{array}{ll} \text{(C)} & E_i\phi \wedge E_i\psi \rightarrow E_i(\phi \wedge \psi) \\ \text{(4)} & E_i\phi \rightarrow E_iE_i\phi \end{array} \quad \begin{array}{ll} \text{(T)} & E_i\phi \rightarrow \phi \\ \text{(RE)} & \vdash (\phi \leftrightarrow \psi) \implies \vdash (E_i\phi \leftrightarrow E_i\psi) \end{array}$$

The operator C_i is a KD-type operator and $C_i\phi$ intends to express that the agent a_i commits itself to make ϕ true. It is defined by the following axiomatics:

$$\begin{array}{ll} \text{(K)} & C_i\phi \wedge C_i(\phi \rightarrow \psi) \rightarrow C_i\psi \\ \text{(Nec)} & \vdash \phi \implies \vdash C_i\phi \end{array} \quad \begin{array}{ll} \text{(D)} & C_i\neg\phi \rightarrow \neg C_i\phi \end{array}$$

Given an atom l , and given these operators, an agent a_i is facing three positions: $C_i E_i l$, $C_i \neg E_i l$ and $\neg C_i E_i l \wedge \neg C_i \neg E_i l$ (respectively, the agent commits itself to make l true, i.e., the agent commits to make an action that makes l true, the agent commits itself not to make l true i.e, the agent commits itself to make no action that will make l true, and the agent does not commit itself to make l true nor commits itself not to make it true).

In this present paper, we forget this axiomatics and we only consider the three sets of atoms: Com_{+,a_i} , which corresponds to $\{l : C_i E_i l\}$, Com_{-,a_i} , which corresponds to $\{l : C_i \neg E_i l\}$, and P_i , which corresponds to $\{l : \neg C_i E_i l \wedge \neg C_i \neg E_i l\}$. But we can check that, by the previous axiomatics, we can derive, as a theorem, $\neg(C_i E_i l \wedge C_i \neg E_i l)$. This explains constraint 1. We can also derive, as a theorem, $\neg(C_i \neg E_i l \wedge C_i E_i l)$. This explains constraint 2.

For defining individual obligations, we only need to consider the positive commitments. So, let us define:

Definition 11.

$$Com_{+,A} = \bigcup_{a_i \in A} Com_{+,a_i}$$

By this definition, $Com_{+,A}$ is composed by any atom an agent commits itself to make true.

Assumption 3 In the following, $Com_{+,A}$ is assumed to be consistent.

This constraint is imposed in order to avoid the case when one agent commits itself to make an atom a true, while another agent commits itself to make that atom false.

3.3 Individual obligations

We can now characterize the obligations that are directed to some agents of the group, given the obligations of the group and given the agent's commitments. Individual obligations are defined by:

Definition 12. Let ϕ be a proposition such that $O_A \phi$. Let a_i be an agent who controls ϕ . We say that a_i is **obligated to satisfy ϕ towards A_ϕ** iff $\models Com_{+,a_i} \rightarrow \phi$. It is denoted $O_{a_i}^{A_\phi} \phi$.

In other words, if the whole group A has the obligation to make ϕ true, (thus if the sub group A_ϕ has the obligation towards A to make ϕ true) and if an agent a_i , belonging to A_ϕ commits itself to achieve ϕ , then it has the individual obligation towards A_ϕ to make ϕ true. This intuitively represents

the fact that, since the sub-group A_ϕ has the obligation to make ϕ true and since a_i , one member of A_ϕ , commits itself towards the other members of A_ϕ to make ϕ true, then it has now the obligation, towards A_ϕ to make ϕ true.

3.4 Satisfaction and violations

For checking if the different obligations introduced previously are violated or not, we must examine the results of the agents' actions.

Let KB_{next} be the state of the world resulting from the actions of the agents.

Let ϕ such that $O_A\phi$.

- If $KB_{next} \models \phi$ then the collective obligation is not violated.
- If $KB_{next} \not\models \phi$ then $O_A\phi$ is violated.

The whole group A is taken as responsible of the violation, by the agent who directed the collective obligation.

We consider A_ϕ . Since we have $O_A\phi$ we also have $O_{A_\phi}^A\phi$. Thus, since $KB_{next} \not\models \phi$, this proves that $O_{A_\phi}^A\phi$ is violated too. And A_ϕ is taken as responsible, by A , of this violation.

Let us consider all the agents a_i belonging to A_ϕ who committed to achieve ϕ . We thus have $O_{a_i}^{O_\phi}\phi$.

Since, $KB \not\models \phi$, the obligation $O_{a_i}^{O_\phi}\phi$ is violated too and a_i can be taken as responsible, by O_ϕ of this violation.

4 Study of an example

In this section, we will illustrate the previous definitions by an example. Let us consider a group \mathcal{A} of three agents named a_1 , a_2 and a_3 . \mathcal{A} is a group of carpenters which build the doors of a new house. There is a regulation (emitted for instance by the promoter of the building) which imposes the following rules for \mathcal{A} :

- if the door is sanded, then the door should be lacquered and not covered with paper.

- if the door is not sanded, then the door should be covered with paper and not lacquered.

Let us denote by s the fact “the door is sanded”, by p the fact “the door is covered with paper” and by l the fact “the door is lacquered”. The previous scenario is translated into the following set of CO^* formulas : $\{I(l \wedge \neg p | s), I(\neg l \wedge p | s)\}$.

Let us examine some scenarios :

1. let us suppose that $KB = \{s, \neg l, \neg p\}$. The door is sanded, but it is neither lacquered nor covered with paper. Let us also suppose that $C_{a_1} = C_{a_3} = \{l\}$ (i.e. a_1 and a_3 can lacquer the door) and that $C_{a_2} = \{p\}$ (i.e. only a_2 can cover the door with paper). So \mathcal{A} controls both l and p .

In this case, $UI(KB) = \{s\}$ and \mathcal{A} has the obligation of $l \wedge \neg p$. Moreover, as l is controllable by both a_1 and a_3 , then $\{a_1, a_3\}$ has the obligation towards \mathcal{A} to achieve l . Finally, as a_2 is the only agent which controls p , $\{a_2\}$ has the obligation towards \mathcal{A} to achieve $\neg p$.

Thus the obligations are : $O_{\mathcal{A}}(l \wedge \neg p)$, $O_{\{a_1, a_3\}}^{\mathcal{A}}(l)$ and $O_{\{a_2\}}^{\mathcal{A}}(\neg p)$.

- (a) let us suppose that the agents do not commit themselves to anything. Let us also suppose that a_1 , a_2 and a_3 do nothing. In this case, $KB_{next} = KB = \{s, \neg l, \neg p\}$.

As l is a part of the obligation $O_{\mathcal{A}}(l \wedge \neg p)$, the collective obligation is then violated. \mathcal{A} is taken as responsible of this violation.

Moreover, as $\{a_1, a_3\}$ should have lacquered the door ($O_{\{a_1, a_3\}}^{\mathcal{A}}(l)$), $\{a_1, a_3\}$ is taken as responsible by \mathcal{A} of the violation of $O_{\mathcal{A}}(l \wedge \neg p)$.

- (b) let us suppose that the agents do not commit themselves to anything. Let us also suppose that a_1 lacquers the door and that a_2 and a_3 do nothing.

In this case, $KB_{next} = \{s, l, \neg p\}$ and $KB_{next} \models l \wedge \neg p$. The collective obligation imposed on \mathcal{A} is not violated.

- (c) let us suppose that a_1 commits itself to lacquer the door. In this case, $Com_{+, a_1} = \{l\}$ and we can derive $O_{a_1}^{\{a_1, a_3\}}(l)$. a_1 is obligated to achieve l towards $\{a_1, a_3\}$.

Assume that a_1 lacquers the door and that a_2 and a_3 do nothing. In this case, $KB_{next} = \{s, l, \neg p\}$ and all the obligations are not violated.

Assume now that a_1 does not lacquer the door, but that a_3 lacquers the door. In this case, the collective obligation $O_{\mathcal{A}}(l \wedge \neg p)$ is satisfied, $O_{\{a_1, a_3\}}^{\mathcal{A}}(l)$ is satisfied too, but $O_{a_1}^{\{a_1, a_3\}}(l)$ is violated. Even if the group fulfilled its obligations, the obligation of a_1 towards $\{a_1, a_3\}$ to achieve l is violated.

- (d) let us suppose that a_1 commits itself to lacquer the door. In this case, $Com_{+, a_1} = \{l\}$ and we can derive $O_{a_1}^{\{a_1, a_3\}}(l)$. a_1 is obligated to achieve l towards $\{a_1, a_3\}$.

Let us suppose that a_1 and a_3 do nothing and that a_2 covers the door with paper. In this case, as $KB_{next} = \{s, \neg l, p\}$, the collective obligation for the group is violated. a_2 has also violated its obligation toward the group \mathcal{A} to do $\neg p$. Finally, $\{a_1, a_3\}$ has violated its obligation to do l toward \mathcal{A} and a_1 has violated its obligation to do l toward $\{a_1, a_3\}$.

2. let us now suppose that $KB = \{\neg s, \neg l, \neg p\}$, i.e. the door is neither sanded, nor lacquered nor covered with paper. Let us also suppose that $C_{a_1} = \{l\}$, $C_{a_2} = \{p\}$ and $C_{a_3} = \{s\}$. Thus \mathcal{A} controls l , p and s .

As $UI(KB) = \phi$, there are two obligations for \mathcal{A} : $O_{\mathcal{A}}(s \rightarrow l \wedge \neg p)$ and $O_{\mathcal{A}}(\neg s \rightarrow \neg l \wedge p)$.

Let us suppose that a_3 commits itself to sand the door. As $(s \rightarrow l \wedge \neg p)$ and $(\neg s \rightarrow \neg l \wedge p)$ are controllable only by $\{a_1, a_2, a_3\}$, we cannot derive any other obligation. If a_3 does not sand the door, a_1 lacquers the door and a_2 does nothing (i.e. if $KB_{next} = \{\neg s, l, \neg p\}$, then $O_{\mathcal{A}}(\neg s \rightarrow \neg l \wedge p)$ is violated by \mathcal{A} . That is the only violation we can derive. Intuitively, it would have been correct to derive that a_3 has an obligation to do s with respect to \mathcal{A} . a_1 and a_2 have acted as if a_3 respected its commitments and are not responsible, toward the group, for the violation of the collective obligation.

To be able to derive such individual obligations, we could extend definition 12 to :

Let ϕ be a proposition such that $O_{\mathcal{A}}(\phi)$. Let a_i be an agent in \mathcal{A}_{ϕ} . If $\models Com_{+, a_i} \rightarrow \gamma$ and $\{\gamma\} \cup \phi$ is consistent, then a_i is obligated to

achieve γ towards \mathcal{A}_ϕ .

Unfortunately, with this definition, if the agent commits itself to close the window for instance (which can be modeled by w), then we can derive $O_{a_3}^A(w)$, even if closing the window has no link with the collective obligation imposed on the group.

5 Discussion

In this paper, we have presented a very preliminary work about collective obligations, i.e. obligations directed to a group of agents.

We have assumed that there was no hierarchical structure in the group, and no institutionalized agent who represents the group like in [4]: the group is made of real agents who may coordinate or not to act on the world.

In this work, the collective obligations are represented by conditional preferences. The first step was to determine the obligations of the group, given what is fixed in the world and given what this group as a whole, can do. Then we considered that, if the group is obliged to make A true, then it induces another obligation to the very sub-group who control A : that sub-group is obliged, towards the whole group, to make A true. These definitions of obligation are direct extensions, to the multi-agent case, of one definition provided by Boutilier in the single-agent case.

As for individual obligations, they are induced as soon as an agent commits itself to satisfy, by one of its action, an obligation of the group. Checking if these obligations are violated or not need to consider the state of the world obtained after the agents' actual actions.

The study of an example shows that the definitions given in the paper are rather encouraging but need to be refined as the case 2 in section 4 shown it. Furthermore, this work could be extended in many directions.

For instance, concerning the agent's model, it would be interesting to relate the notion of commitment used here with the notion of proposition which are "controllable and fixed" defined in [5].

Secondly, one must notice that the notion of controllability taken here has an important weakness: if l is controllable, then $\neg l$ is also controllable. This is questionable since having the ability to make an atom true does not necessarily mean having the ability to make its negation true. For instance, even if one is able to sand a door down, it is not necessarily able to "unsand" it down.

We are currently working on a more refined model of ability in which an agent may control an atom but not its negation. In this refined model, we also intend to take into account the fact that some atoms are controllable not by a single agents but by a coalition of agents [9]: for instance, two agents are needed to raise a door. The impact of this refinement to the previous work remains to be studied.

Acknowledgements We would like to thank the anonymous referee who carefully read this paper and whose remarks helped us to improve this work.

References

- [1] N. Belnap and M. Perloff. Seeing to it that: a canonical for of agencies. *Theoria*, 54:175–199, 1988.
- [2] C. Boutilier. Conditional logics of normality: a modal approach. *Artificial Intelligence*, 68:87–154, 1994.
- [3] C. Boutilier. Toward a logic for qualitative decision theory. In *Principles of Knowledge representation and Reasoning (KR'94)*. J. Doyle, E. Sandewall and P. Torasso Editors, 1994.
- [4] J. Carmo and O. Pacheco. Deontic logic and action logics for organized collective agency. *Fundamenta Informaticae*, 48:129–163, 2001.
- [5] L. Cholvy and Ch. Garion. An attempt to adapt a logic of conditional preferences for reasoning with contrary-to-duties. *Fundamenta Informaticae*, 48:183–204, 2001.
- [6] Ch. Garion. Distributions des exigences : Un problème de calcul de buts individuels en fonction de buts collectifs. In M. Ayel and J.-M. Fouet, editors, *Actes des Cinquièmes Rencontres Nationales des Jeunes Chercheurs en Intelligence Artificielle*, 2000. *In french*.
- [7] 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.
- [8] L. Royakkers and F. Dignum. No organization without obligations: How to formalize collective obligation? In M. Ibrahim, J. Kung, and N. Revell, editors, *Proceedings of 11th International Conference on Databases and Expert Systems Applications (LNCS-1873)*,, pages 302–311. Springer-Verlag, 2000.
- [9] O. Shehory and S. Kraus. Methods for task allocation via agent coalition formation. *Artificial Intelligence*, 101(1-2):165–200, 1998.