

# Deriving individual obligations from collective obligations

Christophe Garion<sup>1</sup>, Laurence Cholvy<sup>2</sup>

<sup>1</sup> SUPAERO

10 av. Édouard Belin

31055 Toulouse, France [garion@supaero.fr](mailto:garion@supaero.fr)

<sup>2</sup> ONERA-Toulouse

2 bis av. Édouard Belin

31055 Toulouse, France [cholvy@cert.fr](mailto:cholvy@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 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.

**Keywords.**

## 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 [1].

According to Royakkers and Dignum [2], a collective obligation is an obligation directed to a group of individuals i.e. a group of agents. For instance (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.

Norman and Reed [3] use the terms *collective group* and *distributive group* to make this distinction. If distributive, a group is addressed distributively (“*Boys, you have to eat properly*”); if collective, a group is being addressed as a collective (“*Boys, you have to set the table*”).

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. Notice that the collective obligation will be violated too if the eldest one, who is the tallest and the only one who can take the glasses, does not take the glasses, even if the two youngest boys carry the forks, the knives and the plates. As said previously, in this case 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 [4], [5], 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 [4], 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 (cf. also [6]). 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  $\bar{\square}$ . Models of

$CO$  are of the form:  $M = \langle W, \leq, \phi \rangle$  where  $W$  is a set of worlds,  $\phi$  is a valuation function<sup>1</sup>, and  $\leq$  is a total pre-order<sup>2</sup> on worlds, allowing one to express preference:  $v \leq w$  means that  $v$  is at least as preferred as  $w$ .

**Definition 1.** Let  $M = \langle W, \leq, \phi \rangle$  be a  $CO$  model. The valuation of a formula in  $M$  is given by:

$$\begin{aligned}
M \models_w \alpha & \quad \text{iff } w \in \phi(\alpha) \text{ for any propositional} \\
& \quad \text{letter } \alpha. \\
M \models_w \neg\alpha & \quad \text{iff } M \not\models_w \alpha \text{ for any formula } \alpha. \\
M \models_w (\alpha_1 \wedge \alpha_2) & \quad \text{iff } M \models_w \alpha_1 \text{ and } M \models_w \alpha_2 \text{ if } \alpha_1 \\
& \quad \text{and } \alpha_2 \text{ are formulas.} \\
M \models_w \Box\alpha & \quad \text{iff for any world } v \text{ such that } v \leq w, \\
& \quad M \models_v \alpha \\
M \models_w \bar{\Box}\alpha & \quad \text{iff for any world } v \text{ such that } w < v, \\
& \quad M \models_v \alpha \\
M \models \alpha & \quad \text{iff } \forall w \in W \ M \models_w \alpha.
\end{aligned}$$

Thus,  $\Box\alpha$  is true in the world  $w$  iff  $\alpha$  is true in all the worlds which are at least as preferred as  $w$ .  $\bar{\Box}\alpha$  is true in the world  $w$  iff  $\alpha$  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  $\bar{\Diamond}\alpha \equiv_{def} \neg\bar{\Box}\neg\alpha$ . Furthermore, Boutilier defines:  $\bar{\Box}\alpha \equiv_{def} \Box\alpha \wedge \bar{\Box}\alpha$  and  $\bar{\Diamond}\alpha \equiv_{def} \Diamond\alpha \vee \bar{\Diamond}\alpha$ .

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

Boutilier then considers  $CO^*$  [5], 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:  $M \models \bar{\Diamond}A$ , 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:

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

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

An absolute preference is of the form:  $I(A|\top)$ <sup>3</sup>. 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:

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

<sup>2</sup>  $\leq$  is reflexive, transitive and connected binary relation.

<sup>3</sup> Where  $\top$  is any propositional tautology.

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

*Example 1.* Suppose an employee who is requested to make sure that the proposal for the financial management of the next year is written unless the statistics are not collated. We consider a propositional language whose letters are  $s$  (the statistics for the current year are collated) et  $fp$  (the proposal for the financial management of the next year is written) and we consider the two conditionals  $I(fp)$  and  $I(\neg fp|\neg s)$  which express that the proposal for the financial management for the next year should be written, but if the statistics of the current year are not collated, then it is preferred that it this proposal is not written. The possible worlds are:  $w_1 = \{fp, s\}$ ,  $w_2 = \{\neg fp, \neg s\}$ ,  $w_3 = \{fp, \neg s\}$ ,  $w_4 = \{\neg fp, s\}$ .<sup>5</sup> Because of  $I(fp)$ , the worlds  $w_1$  and  $w_3$  can be the most preferred ones. But, due to  $I(\neg fp|\neg s)$ ,  $w_3$  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  $I(\neg fp|\neg s)$ . Thus  $w_2 \leq w_3$ . The models which satisfy  $I(fp)$  and  $I(\neg fp|\neg s)$  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 statistics are collated). Thus  $Cl(KB_1) = \{s\}$ . Ideal goals for the agent are  $\alpha$  such that  $\forall M M \models I(\alpha|s)$ .  $fp$  is thus an ideal goal for the agent: since the statistics are collated, the agent has to write the financial proposal. Assume now that  $KB_2 = \{\neg s\}$  (the statistics are not collated). One can prove that  $\neg fp$  is now the ideal goal of the agent: since the statistics are not collated, the agent must not write the financial proposal. This is questionable and discussed in the following section.

Notice that this example is inspired from a scenario studied in [3]. Here, we modified this scenario in order to get conditional obligations and we restrict it to a single agent. Moreover, an extension of this scenario will be studied in section 4.

## 2.2 Controllable, influenceable propositions and CK-goals

By definition 2, any formula  $\alpha$  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,

<sup>4</sup> In 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.

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

in the second case of example 1, if the agent can collect statistics, it would be preferable that he does so, and that he also write the proposal 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.

Furthermore, it may happen that some formulas  $\alpha$  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 collect the statistics (for instance, s/he does not have permission to access the database): collecting statistics cannot be a goal for the agent.

So, Boutilier introduces a partition of atoms of  $PROP$ :  $PROP = C \cup \bar{C}$ .  $C$  is the set of the atoms the agent controls (i.e. the atoms the agent can change the truth value) and  $\bar{C}$  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 the access to the appropriate database and he can use the dedicated software, 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  $\bar{C}$  respectively be the set of atoms that the agent controls and the set of atoms that he does not control. A proposition  $\alpha$  is *controllable* iff, for any  $u \in V(\bar{C})$ , 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  $\alpha$  is *influenceable* iff there are  $u \in V(\bar{C})$ ,  $v \in V(C)$  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 propositions 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.<sup>6</sup> Under this assumption, Boutilier then defines the notion of CK-goal:

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

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

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:

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

*Example 2.* Consider again  $\Sigma = \{I(fp), I(\neg fp|\neg s)\}$ . Assume that  $KB = \{\neg fp, \neg s\}$  (the statistics are not collected and the financial proposal is not written). Assume first that the agent can collect the statistics and also write the financial proposal. Then  $UI(KB) = \emptyset$ . Thus,  $\{fp, s\}$  is the atomic goal set of the agent: the agent has to collect the statistics and to write the proposal. Assume now that the agent can collect the statistics but cannot write the financial proposal. Here,  $fp$  is not controllable, thus  $UI(KB) = \{\neg fp\}$  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 1** *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 8.** Let  $C = \bigcup_{a_i \in \mathcal{A}} C(a_i)$  and  $\overline{C} = PROP \setminus C$ . A proposition  $\alpha$  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  $\alpha$  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 3.* 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 [6], 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 4.* 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 9.** Given a knowledge-base  $KB$ , a proposition  $\alpha$  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 10.** 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 11.** The group  $\mathcal{A}$  has the obligation of  $\varphi$  towards the agent who directed the collective obligation iff  $\Sigma \models I(\varphi|UI(KB))$  with  $\varphi$  controllable by  $\mathcal{A}$ . It is denoted  $O_{\mathcal{A}}\varphi$ .

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 [6].

Thus, these obligations characterize the most preferred situation that the group  $\mathcal{A}$  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 12.** *Let  $\phi$  be a proposition. Let  $\mathcal{A}_\phi$  be the union of the minimal subsets of  $\mathcal{A}$  which controls  $\phi$ <sup>7</sup>. We say that **the sub-group  $\mathcal{A}_\phi$  has the obligation of  $\phi$ , towards  $\mathcal{A}$**  iff  $\Sigma \models I(\phi|UI(KB))$ . It is denoted  $O_{\mathcal{A}_\phi}^{\mathcal{A}}\phi$ .*

Thus, these obligations characterize the most preferred situation that the group  $\mathcal{A}_\phi$  (a group which is the union of all the minimal sub-groups which control  $\phi$ ) can achieve, given what is fixed.

*Example 5.* Let us now extend the example 1 to a group of agents and consider three agents *Alice*, *John* and *Tom* who are requested to write a financial proposal a scientific proposal. The conditional preference which models this collective obligation is:  $I(fp \wedge sf)$ .

Assume that *Alice* is able to write the financial proposal while *John* and *Tom* are able to write the scientific proposal. Then we have:

$$\begin{aligned} &O_{\{Alice, John, Tom\}}(fp \wedge sp) \\ &O_{\{Alice\}}^{\{Alice, John, Tom\}}fp \\ &O_{\{John, Tom\}}^{\{Alice, John, Tom\}}sp \end{aligned}$$

In other words, the group  $\{Alice, John, Tom\}$  has the obligation to write the two proposals. The singleton  $\{Alice\}$  has the obligation, towards the whole group to write the financial proposal, and the sub-group  $\{John, Tom\}$  has the obligation, towards the whole group to write the scientific proposal.

### 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  $a_i$

<sup>7</sup> We could also choose to define  $\mathcal{A}_\phi$  as some of the minimal subsets of  $\mathcal{A}$  which controls  $\phi$ . However, the whole study of the consequence of this alternative has not yet been done.

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**  $\forall a_i \in \mathcal{A} \quad Com_{+,a_i}$  is consistent.

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

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 1.* The previous notions have been modeled in modal logic in [7], with two families of modal operators:  $C_i$  and  $E_i$ ,  $i \in \{1 \dots n\}$ . The operator  $E_i$  is the *stit* operator ([8], [9]).  $E_i\phi$  intends to express that the agent  $a_i$  is seeing to it that  $\phi$ . It is defined by the following axiomatics:

- (C)  $E_i\phi \wedge E_i\psi \rightarrow E_i(\phi \wedge \psi)$       (T)  $E_i\phi \rightarrow \phi$   
(4)  $E_i\phi \rightarrow E_iE_i\phi$       (RE)  $\vdash (\phi \leftrightarrow \psi) \implies \vdash (E_i\phi \leftrightarrow E_i\psi)$

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  $\phi$  true. It is defined by the following axiomatics:

- (K)  $C_i\phi \wedge C_i(\phi \rightarrow \psi) \rightarrow C_i\psi$       (D)  $C_i\neg\phi \rightarrow \neg C_i\phi$   
(Nec)  $\vdash \phi \implies \vdash C_i\phi$

Given an atom  $l$ , and given these operators, an agent  $a_i$  is facing three positions:  $C_iE_i l$ ,  $C_i\neg E_i l$  and  $\neg C_iE_i l \wedge \neg C_i\neg E_i l$  (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  $\{l : C_iE_i l\}$ ,  $Com_{-,a_i}$ , which corresponds to  $\{l : C_i\neg E_i l\}$ , and  $P_i$ , which corresponds to  $\{l : \neg C_iE_i l \wedge \neg C_i\neg E_i l\}$ . But we can check that, by the previous axiomatics, we can derive, as a theorem,  $\neg(C_iE_i l \wedge C_i\neg E_i l)$ . This explains constraint 1. We can also derive, as a theorem,  $\neg(C_i\neg E_i l \wedge C_iE_i l)$ . This explains constraint 2.

For defining individual obligations, we only need to consider the positive commitments (this assumption will be discussed in section 5). So, let us define:

**Definition 13.**

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

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

**Assumption 2** *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 14.** *Let  $\phi$  be a proposition such that  $O_A\phi$  holds. Let  $a_i$  be an agent of  $\mathcal{A}$ . If there is some minimal  $\{l_1, \dots, l_m\} \subseteq Com_{+,a_i}$  such that  $\models l_1 \wedge \dots \wedge l_m \rightarrow \phi$ , we say that  $a_i$  is **obligated to satisfy  $l_1 \wedge \dots \wedge l_m$  towards  $\mathcal{A}_\phi$** . This is denoted by  $O_{a_i}^{A_\phi}(l_1 \wedge \dots \wedge l_m)$ .*

Several remarks can be done on this definition. Let us suppose that the whole group  $\mathcal{A}$  has the obligation to make  $\phi$  true, (thus the sub group  $\mathcal{A}_\phi$  has the obligation towards  $\mathcal{A}$  to make  $\phi$  true). Let us suppose that there is an agent  $a_i$  such that there is some minimal  $\{l_1, \dots, l_m\} \subseteq Com_{+,a_i}$  such that  $\models l_1 \wedge \dots \wedge l_m \rightarrow \phi$ .

First, the set  $\{l_1, \dots, l_m\}$  is said to be minimal, in the sense that for any  $j \in \{1, \dots, m\}$   $(\{l_1, \dots, l_m\} - \{l_j\}) \not\models \phi$ .  $\{l_1, \dots, l_m\}$  can be viewed as a kind of “prime implicant” of  $\phi$ , because we do not want to derive individual obligations with no link with the obligation imposed to the group. For instance, we can have  $O_A(fp)$ : the group have the obligation to do a financial proposal. Let us suppose that *Tom* commits itself to do the financial proposal and to wash its car. In this case, with no minimality condition, we could derive that *Tom* is obligated to do the financial proposal and to wash its car towards the sub-group of agents “reponsible” for doing the financial proposal. But the washing of *Tom*'s car has no link with the collective obligation of doing a financial proposal. The minimality condition ensures that the individual obligations will not be “out of context”.

If an agent  $a_i$  commits itself to achieve some “actions”  $l_1, \dots, l_m$  such that  $\models l_1 \wedge \dots \wedge l_m \rightarrow \phi$ , then it has the individual obligation towards  $\mathcal{A}_\phi$  to make  $l_1 \wedge \dots \wedge l_m$  true (notice that in [10], the agent has the individual obligation to make  $\phi$  true. The new formulation of individual obligation is more precise). This intuitively represents the fact that, since the sub-group  $\mathcal{A}_\phi$  has the obligation to make  $\phi$  true and since  $a_i$  commits itself towards the other members of  $\mathcal{A}_\phi$  to make some sufficient conditions of  $\phi$  true, then it has now the obligation, towards  $\mathcal{A}_\phi$  to make those sufficient conditions true.

Finally, let us remark that such an agent  $a_i$  belongs to  $\mathcal{A}_\phi$ , because it controls some literals (in fact  $l_1, \dots, l_m$ ) whose conjunction implies  $\phi$ .

### 3.4 Satisfactions 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  $\phi$  such that  $O_{\mathcal{A}}\phi$ .

- if  $KB_{next} \models \phi$  then the collective obligation is not violated. We say that the collective obligation is fulfilled.
- if  $KB_{next} \not\models \phi$  then  $O_{\mathcal{A}}(\phi)$  is violated.  
The whole group  $\mathcal{A}$  is taken as responsible of the violation, by the agent who directed the collective obligation.  
We consider  $\mathcal{A}_\phi$ . Since we have  $O_{\mathcal{A}}\phi$  we also have  $O_{\mathcal{A}_\phi}^{\mathcal{A}}(\phi)$ . Thus, since  $KB_{next} \not\models \phi$ , this proves that  $O_{\mathcal{A}_\phi}^{\mathcal{A}}(\phi)$  is violated too. And  $\mathcal{A}_\phi$  is taken as responsible, by  $\mathcal{A}$ , of this violation.
- let us consider all the agents  $a_i$  such that there is some  $\varphi$  such that  $O_{a_i}^{\mathcal{A}_\phi}(\varphi)$ .  
If  $KB_{next} \not\models \varphi$ , the obligation  $O_{a_i}^{\mathcal{A}_\phi}(\varphi)$  is violated too and  $a_i$  can be taken as responsible by  $\mathcal{A}_\phi$  of the violation of its commitment i.e.  $O_{a_i}^{\mathcal{A}_\phi}\varphi$ .  
Moreover, if  $KB_{next} \not\models \phi$ ,  $a_i$  can be taken by  $\mathcal{A}_\phi$  of the violation of  $O_{\mathcal{A}_\phi}^{\mathcal{A}}(\phi)$ .

## 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 Alice (denoted by  $A$ ), John (denoted by  $J$ ) and Tom (denoted by  $T$ ). That group is addressed the following obligations:

- if the statistics are collected, then the financial proposal and the scientific proposal should be written.
- if the statistics are not collected, then the financial proposal should not be written, but the scientific proposal should be.

Let us denote by  $s$  the fact “the statistics are collected”, by  $fp$  the fact “the financial proposal is written” and by  $sp$  the fact “the scientific proposal is written”. The previous scenario is translated into the following set of  $CO^*$  formulas:  $\{I(fp \wedge sp|s), I(\neg fp \wedge sp|\neg s)\}$ .

Let us examine some scenarios:

1. let us suppose that  $KB = \{s, \neg fp, \neg sp\}$ . The statistics are collected, but neither the financial proposal nor the scientific proposal are written. Let us also suppose that  $C_A = C_T = \{fp\}$  (i.e. Alice and Tom can write the financial proposal) and that  $C_J = \{sp\}$  (i.e. only John can write the scientific proposal). So  $\mathcal{A}$  controls both  $fp$  and  $sp$ .  
In this case,  $UI(KB) = \{s\}$  and  $\mathcal{A}$  has the obligation of  $fp \wedge sp$ , thus  $\mathcal{A}$  has the obligation of  $fp$  and the obligation of  $sp$ . Moreover, as  $fp$  is controllable by both Alice and Tom, then  $\{A, T\}$  has the obligation towards  $\mathcal{A}$  to achieve  $fp$ . Finally, as John is the only agent which controls  $sp$ ,  $\{J\}$  has the obligation towards  $\mathcal{A}$  to achieve  $sp$ .  
Thus the obligations are :  $O_{\mathcal{A}}(fp \wedge sp)$ ,  $O_{\{A, T\}}^{\mathcal{A}}(fp)$  and  $O_{\{J\}}^{\mathcal{A}}(sp)$ .

- (a) let us suppose that the agents do not commit themselves to anything. Let us also suppose that Alice, John and Tom do nothing. In this case,  $KB_{next} = KB = \{s, \neg fp, \neg sp\}$ . As  $fp$  and  $sp$  are parts of the obligation  $O_{\mathcal{A}}(fp \wedge sp)$ , the collective obligation is then violated.  $\mathcal{A}$  is taken as responsible of this violation. Moreover, as  $\{A, T\}$  should have written the financial proposal ( $O_{\{A, T\}}^{\mathcal{A}}(fp)$ ),  $\{A, T\}$  is taken as responsible by  $\mathcal{A}$  of the violation of  $O_{\mathcal{A}}(fp)$ . By the same way,  $\{J\}$  is taken as responsible of the violation of  $O_{\mathcal{A}}(sp)$  by  $\mathcal{A}$ .
- (b) let us suppose that the agents do not commit themselves to anything. Let us also suppose that Alice writes the financial proposal and that John and Tom do nothing. In this case,  $KB_{next} = \{s, fp, \neg sp\}$  and  $KB_{next} \models fp \wedge \neg sp$ . The collective obligation imposed on  $\mathcal{A}$  is violated and the group is taken as responsible of the violation of  $fp \wedge sp$ . More precisely,  $O_{\mathcal{A}}(fp)$  is fulfilled because Alice wrote the financial proposal. But  $O_{\mathcal{A}}(sp)$  is violated. As previously,  $\{J\}$  is taken as responsible of the violation of  $O_{\mathcal{A}}(sp)$  by  $\mathcal{A}$ .
- (c) let us suppose that Alice commits herself to write the financial proposal. In this case,  $Com_{+, \mathcal{A}} = \{fp\}$  and we can derive  $O_{\mathcal{A}}^{\{A, T\}}(fp)$  (because  $O_{\mathcal{A}}(fp)$  holds). Alice is obligated to achieve  $fp$  towards  $\{A, T\}$ . Assume that Alice writes the financial proposal, that John writes the scientific proposal and that Tom does nothing. In this case,  $KB_{next} = \{s, sp, fp\}$  and all the obligations are fulfilled. Assume now that Alice does not write the financial proposal, but that Tom writes the financial proposal. Assume also that John writes the scientific proposal. In this case, the collective obligation  $O_{\mathcal{A}}(fp \wedge sp)$  is satisfied,  $O_{\{A, T\}}^{\mathcal{A}}(fp)$  is satisfied too, but  $O_{\mathcal{A}}^{\{A, T\}}(fp)$  is violated. Even if the group fulfilled its obligations, the obligation of Alice towards  $\{A, T\}$  to achieve  $fp$  is violated. Let us finally suppose that Alice, John and Tom do nothing. In this case, as  $KB_{next} = \{s, \neg sp, \neg fp\}$ , the collective obligation for the group is violated. John has also violated its obligation toward the group  $\mathcal{A}$  to do  $sp$ . Finally,  $\{A, T\}$  has violated its obligation to do  $fp$  toward  $\mathcal{A}$  and Alice has violated its obligation to do  $fp$  toward  $\{A, T\}$ .
2. let us now suppose that  $KB = \{\neg s, \neg fp, \neg sp\}$ , i.e. the statistics are not collected and neither the financial proposal nor the scientific proposal are written. Let us also suppose that  $C_A = \{fp\}$ ,  $C_J = \{sp\}$  and  $C_T = \{s\}$ . Thus  $\mathcal{A}$  controls  $l$ ,  $p$  and  $s$ . As  $UI(KB) = \phi$ , there are three obligations for  $\mathcal{A}$ :  $O_{\mathcal{A}}(s \rightarrow fp)$ ,  $O_{\mathcal{A}}(\neg s \rightarrow \neg fp)$  and  $O_{\mathcal{A}}(sp)$ . Thus we have:  $O_{\{J\}}^{\mathcal{A}}(sp)$ ,  $O_{\{A, T\}}^{\mathcal{A}}(s \rightarrow fp)$  and  $O_{\{A, T\}}^{\mathcal{A}}(\neg s \rightarrow \neg fp)$ . Let us suppose that Tom commits himself to collect the statistics. As  $\models s \rightarrow (\neg s \rightarrow \neg fp)$ , then Tom has the obligation towards  $\{A, T\}$  to do  $s$ , i.e.,  $O_{\{T\}}^{\{A, T\}}s$ .

Suppose now that Alice wrote the financial proposal (because she thought that Tom would collect the statistics) but that Tom does not collect them. Suppose also that John does nothing. Then,  $KB_{next} = \{\neg s, fp, \neg sp\}$ . In this case,  $O_{\mathcal{A}}(\neg s \rightarrow \neg fp)$  and  $O_{\mathcal{A}}(sp)$  are both violated by  $\mathcal{A}$ . John violated also his obligation  $O_{\{J\}}^{\mathcal{A}}sp$  and  $\{A, T\}$  violated its obligation  $O_{\{A, T\}}^{\mathcal{A}}(\neg s \rightarrow \neg fp)$ . But Tom can be taken as responsible by  $\{A, T\}$  (and in particular by Alice) of the violation of  $O_{\{A, T\}}^{\mathcal{A}}(\neg s \rightarrow \neg fp)$ , because  $O_{\{T\}}^{\{A, T\}}s$  holds.

## 5 Discussion

In this paper, we have presented a 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 [1]: 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.

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 [6]. We could also refine the notion of commitment to the notion of commitment toward a group of agent. For instance, Tom can commit himself to write the financial proposal toward Alice and John et commit himself to wash the car toward his wife. Using this distinction could refine the obligations derivation process. Moreover, we only consider that the agents commit themselves to make a literal true. We could extend this to propositional formulas. But in this case, the derivation process is more complicated. For instance, if an agent commits him/herself to do  $a \vee b$  and  $a$  and  $b$  are implicants for two different obligations, it is difficult to determine what the other agents should do.

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 can send emails, he/she cannot send emails back once they are sent.

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 [11]: for instance, several agents (in several areas) are needed to collect statistics. The impact of this refinement to the previous work remains to be studied.

Notice also that the agents share the same beliefs about the world. What happens when the group of agents does not share a common set of beliefs ? For instance, Alice may believe that the statistics are not collected and act in this way and Tom may believe that the statistics are collected and write the financial proposal. To solve such conflicts, we could for instance use some kind of merging methods which are used to build a common belief set from several belief sets which can be contradictory [12,13,14]. Particularly, we suppose that the agents beliefs are knowledge, in the sense that what they believe is true in the real world. We could also suppose that the agents beliefs do not fit the actual world. In this case, is an agent responsible of some obligations violations if it has acted as its beliefs were true?

Concerning the definition of individual obligations, we only use the “positive” commitments of the agents. But each agent can also express commitment of the kind “I commit myself not to do the financial proposal”. As there are two sets  $Com_+$  and  $Com_-$  for each agent, there must be two kinds of individual obligations. In our formalism, we express individual obligations to *do* something but no individual obligations *not to do* something.

## References

1. Carmo, J., Pacheco, O.: Deontic and action logics for organized collective agency, modeled through institutionalized agents and roles. *Fundamenta Informaticae* **48** (2001) 129–163
2. Royakkers, L., Dignum, F.: No organization without obligations: how to formalize collective obligation? In Klusch, M., Kerschberg, L., eds.: 11th International Conference on Databases and Expert Systems Applications (LNCS-1873), Springer-Verlag (2000) 191–207
3. Norman, T., Reed, C.: Group delegation and responsibility. In: Proceedings of the first International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS’02), ACM Press (2002) 491–498
4. Boutilier, C.: Toward a logic for qualitative decision theory. In Doyle, J., Sandewall, E., Torasso, P., eds.: Principles of Knowledge Representation and Reasoning (KR’94), Morgan Kaufmann (1994) 75–86
5. Boutilier, C.: Conditional logics of normality : a modal approach. *Artificial Intelligence* **68** (1994) 87–154
6. Cholvy, L., Garion, C.: An attempt to adapt a logic of conditional preferences for reasoning with contrary-to-duties. *Fundamenta Informaticae* **48** (2001) 183–204
7. Cholvy, L., Garion, C.: Strategies for distributing goals in a team of cooperative agents. In Gleizes, M.P., Omicini, A., Zambonelli, F., eds.: Proceedings of the Fifth International Workshop on Engineering Societies in the Agents World (ESAW’04). Number 3451 in Lecture Notes in Artificial Intelligence, Springer-Verlag (2005) 178–190

8. Belnap, N., Perloff, M.: Seeing to it that: a canonical form for agentives. *Theoria* **54** (1988) 175–199
9. Horty, J., Belnap, N.: The deliberative stit : a study of action, omission, ability and obligation. *Journal of Philosophical Logic* **24** (1995) 583–644 Reprinted in *The Philosopher's Annual, Volume 18-1995*, Ridgeview Publishing Company, 1997.
10. Cholvy, L., Garion, C.: Collective obligations, commitments and individual obligations: a preliminary study. In Horty, J., Jones, A., eds.: *Proceedings of the 6<sup>th</sup> International Workshop on Deontic Logic In Computer Science (ΔEON'02)*, Londres (2002) 55–71
11. Kraus, S., Shehory, O.: Methods for task allocation via agent coalition formation. *Artificial Intelligence* **101** (1998) 165–200
12. Konieczny, S., Pino-Pérez, R.: Merging information under constraints: a qualitative framework. *Journal of Logic and Computation* **12** (2002) 773–808
13. Cholvy, L., Garion, C.: Answering queries addressed to merged databases: a query evaluator which implements a majority approach. In Hacid, M.S., Raś, Z., Zighed, D., Kodratoff, Y., eds.: *Foundations of Intelligent Systems - Proceedings of the 13<sup>th</sup> International Symposium on Methodologies for Intelligent Systems, ISMIS 2002*. Volume 2366 of *Lecture Notes in Artificial Intelligence.*, Springer (2002) 131–139
14. Cholvy, L., Garion, C.: Querying several conflicting databases. *Journal of Applied Non-Classical Logics* **3** (2004) 295–327