Formal Handling of Threats and Rewards in a Negotiation Dialogue

Leila Amgoud and Henri Prade

Institut de Recherche en Informatique de Toulouse (I.R.I.T.)–C.N.R.S.
Université Paul Sabatier, 118 route de Narbonne,
31062 Toulouse Cedex 4, France
{amgoud, prade}@irit.fr

Abstract. Argumentation plays a key role in finding a compromise during a negotiation dialogue. It may lead an agent to change its goals/ preferences and force it to respond in a particular way. Two types of arguments are mainly used for that purpose: *threats* and *rewards*. For example, if an agent receives a threat, this agent may accept the offer even if it is not fully "acceptable" for it (because otherwise really important goals would be threatened).

The contribution of this paper is twofold. On the one hand, a logical setting that handles these two types of arguments is provided. More precisely, logical definitions of threats and rewards are proposed together with their weighting systems. These definitions take into account that negotiation dialogues involve not only agents' beliefs (of various strengths), but also their goals (having maybe different priorities), as well as the beliefs about the goals of other agents.

On the other hand, a "simple" protocol for handling such arguments in a negotiation dialogue is given. This protocol shows when such arguments can be presented, how they are handled, and how they lead agents to change their goals and behaviors.

Keywords: Argumentation, Negotiation.

1 Introduction

Negotiation is the predominant interaction mechanism between autonomous agents looking for a compromise. Indeed, agents make offers that they find acceptable and respond to offers made to them.

Recent works on negotiation [2, 3, 4, 6, 7, 9, 10, 11] have argued that argumentation can play a key role in finding the compromise. Indeed, an offer supported by a 'good argument' has a better chance to be accepted, because the argument brings new information possibly ignored by the receiver. If this information conflicts with previous beliefs of the receiver, this agent may even revise its beliefs if it has no strong counter-argument for challenging the information. Moreover, argumentation may constrain the future behavior of the agent, especially if it takes the form of a *threat* or of a *reward*. Such arguments complement more classical arguments, called here *explanatory arguments*, which especially aim at providing reasons for believing in a statement. Even if the interest of using threats and

S. Parsons et al. (Eds.): ArgMAS 2005, LNAI 4049, pp. 88–103, 2006.

rewards in a negotiation dialogue [7, 12] has been emphasized, there has been almost no attempt at modeling and incorporating them in a formal dialogue.

This paper aims at providing a logical setting which handles these two types of arguments, together with explanatory arguments. More precisely, logical definitions of threats and rewards are proposed together with their weighting systems. These definitions take into account that negotiation dialogues involve not only agents' beliefs (of various strengths), but also their goals (having maybe different priorities), as well as the beliefs about the goals of other agents. This paper provides also a "simple" protocol for handling such arguments in a negotiation dialogue. This protocol shows when such arguments can be presented, how they are handled, and how they lead agents to change their goals and behaviors.

The paper is organized as follows: Section 2 introduces the logical language for describing the mental states of the agents. Sections 3, 4 and 5 introduce resp. the explanatory arguments, the threats and rewards. For each type of argument, logical definitions are given together with their weighting systems. Note that the given definitions enable us to distinguish between what the agent finds rewarding (resp. threatening) for it and what it finds rewarding (resp. threatening) for the other agent. In section 6, a general argumentation system which handles the three types of arguments is presented. Section 7 introduces a negotiation protocol which is based on the notions of threats and rewards, and which show when such arguments can be presented, how they are handled by their receivers, and how they lead agents to change their behaviors. The approach is illustrated in section 8 on the example of a negotiation between a boss and a worker. In section 9, we conclude by comparing our proposal with existing works and by presenting some perspectives.

2 The Mental States of the Agents

In what follows, $\mathcal{L}$ denotes a propositional language, $\vdash$ classical inference, and $\equiv$ logical equivalence. We suppose that we have two negotiating agents: P (for proponent) and O (for opponent).

Each agent has got a set $\mathcal{G}$ of *goals* to pursue, a knowledge base $\mathcal{K}$, gathering the information it has about the environment, and a base $\mathcal{GO}$, containing what the agent believes the goals of the other agent are. $\mathcal{K}$ may be pervaded with uncertainty (the beliefs are more or less certain), and the goals in $\mathcal{G}$ and $\mathcal{GO}$ may not have equal priority.

Thus, each base is supposed to be equipped with a complete preordering $\geq$. Relation $a \geq b$ holds iff a is at least as certain (resp. as preferred) as b. For encoding it, we use the set of integers $\{0, 1, \ldots, n\}$ as a linearly ordered scale, where n stands for the highest level of certainty or importance and 0 corresponds to the complete lack of certainty or importance. This means that the base $\mathcal{K}$ is partitioned and stratified into

$$\mathcal{K}_1, \ldots, \mathcal{K}_n (\mathcal{K} = \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_n)$$

such that all beliefs in $\mathcal{K}_i$ have the same certainty level and are more certain than beliefs in $\mathcal{K}_j$ where j < i. Moreover, $\mathcal{K}_0$ is not considered since it gathers

formulas which are totally uncertain, and which are not at all beliefs of the agent. Similarly,

$$\mathcal{GO} = \mathcal{GO}_1 \cup \ldots \cup \mathcal{GO}_n and \mathcal{G} = \mathcal{G}_1 \cup \ldots \cup \mathcal{G}_n$$

such that goals in $\mathcal{GO}_i$ (resp. in $\mathcal{G}_i$) have the same priority and are more important than goals in $\mathcal{GO}_j$ (resp. in $\mathcal{G}_j$ where j < i).

Note that some $\mathcal{K}_i$'s (resp. $\mathcal{G}_i$, $\mathcal{GO}_i$) may be empty if there is no piece of knowledge (resp. goal) corresponding to the level i of certainty (resp. importance).

For the sake of simplicity, in all our examples, we only specify the strata that are not empty. Both beliefs and goals are represented by propositional formulas of the language $\mathcal{L}$. Thus a goal is viewed as a piece of information describing a set of desirable states (corresponding to the models of the associated proposition) one of which should be reached.

3 Explanatory Arguments

Explanations constitute the most common category of arguments. In classical argumentation-based frameworks that can handle inconsistency in knowledge bases, each conclusion is justified by arguments. They represent the reasons to believe in a fact.

3.1 Logical Definition

Such arguments have a *deductive* form. Indeed, from premises, a fact or a goal is entailed. Formally:

Definition 1 (Explanatory argument). *An* explanatory *argument is a pair <H, h> such that:*

1. $H \subseteq \mathcal{K}$,
2. $H \vdash h$,
3. *H is consistent and minimal (for $\subseteq$) among the sets satisfying 1) and 2).*

$\mathcal{A}_e$ will denote the set of all the explanatory arguments that can be constructed from $\mathcal{K}$.

Note that the bases of goals are not considered when constructing such arguments (only based on agent's beliefs) in order to avoid *wishful thinking*.

3.2 Strength of Explanatory Arguments

In [1], it has been argued that arguments may have forces of various strengths. These forces will play two roles:

1. they allow an agent to compare different arguments in order to select the 'best' ones,
2. the forces are useful for determining the acceptable arguments among the conflicting ones.

Different definitions of the force of an argument have been proposed in [1]. Generally, this force of an argument can rely on the beliefs from which it is constructed. Explicit priorities between beliefs, or implicit priorities such as specificity, can be the basis for defining the force of an argument. However, different other aspects can be taken into account when defining the force of explanatory arguments. In particular, the length of the argument (in terms of the number of pieces of knowledge involved) may be considered since the shorter is the explanation, the better it is and the more difficult it is to challenge it (provided that it is based on propositions that are sufficiently certain).

When explicit priorities are given between the beliefs, such as certainty levels, the arguments using more certain beliefs are found stronger than arguments using less certain beliefs. The force of an explanatory argument corresponds to the *certainty level* of the less entrenched belief involved in the argument. In what follows, we consider this view of the force. In the case of stratified bases, the force of an argument corresponds to the smallest number of a stratum met by the support of that argument. Formally:

Definition 2 (Certainty level). *Let* $\mathcal{K} = \mathcal{K}_1 \cup \ldots \cup \mathcal{K}_n$ *be a stratified base, and* $H \subseteq \mathcal{K}$.
The certainty level *of* H, *denoted Level(H)* $= min\{j \mid 1 \leq j \leq n$ *such that* $H_j \neq \emptyset\}$, *where* H_j *denotes* $H \cap \mathcal{K}_j$.

Note that $<H, h>$ is all the stronger as Level(H) has a large value.

Definition 3 (Force of an explanation). *Let* $A = <H, h> \in \mathcal{A}_e$. *The* force *of A is Force(A) = Level(H).*

This definition agrees with the definition of an argument as a minimal set of beliefs supporting a conclusion. Indeed, when any member of this minimal set is seriously challenged, the whole argument collapses. This makes clear that the strength of the least entrenched argument fully mirrors the force of the argument whatever are the strengths of the other components in the minimal set. The forces of arguments make it possible to compare any pair of arguments. Indeed, arguments with a higher force are preferred.

Definition 4 (Comparing explanations). *Let* $A, B \in \mathcal{A}_e$. A *is* preferred *to* B $(A \succ_e B)$ *iff Force(A) > Force(B).*

4 Threats

Threats have a negative flavor and are applied to intend to force an agent to behave in a certain way. Two forms of threats can be distinguished:

i) You should do 'a' otherwise I will do 'b',
ii) You should not do 'a' otherwise I will do 'b'.

The first case occurs when an agent P needs an agent O to do 'a' and O refuses. Then, P threatens O to do 'b' which, according to its beliefs, will have bad consequences for O. Let us consider an example.

Example 1. *Let's consider a mother and her child.*

Mother: You should carry out your school work ('a').
Child: No, I don't want to.
Mother: You should otherwise I will not let you go to the party organized by your friend next week-end ('b').

The second kind of threats occurs when an agent O wants to do some action 'a', which is not acceptable for P. In this case, P threatens that if O insists to do 'a' then it will do 'b' which, according to P's beliefs, will have bad consequences for O. The following example from [7] illustrates this kind of threat.

Example 2

Labor union: We want a wage increase ('a').
Manager: I cannot afford that. If I grant this increase, I will have to lay off some employees ('b'). It will compensate for the higher cost entailed by the increase.

4.1 Logical Definition

In all what follows, we suppose that P presents an argument to O. In a dialogue, each agent plays these two roles in turn. For a threat to be effective, it should be painful for its receiver and conflict with at least one of its goals. A threat is then made up of three parts: the *conclusion* that the agent who makes the threat wants, the *threat* itself and finally the *threatened goal*. Moreover, it has an *abductive* form. Formally:

Definition 5 (Threat). *A* threat *is a triple* $<H, h, \phi>$ *such that:*

1. h *is a propositional formula,*
2. $H \subseteq \mathcal{K}$,
3. $H \cup \{\neg h\} \vdash \neg\phi$ *such that* $\phi \in \mathcal{GO}$,
4. $H \cup \{\neg h\}$ *is consistent and* H *is minimal (for set inclusion) among the sets satisfying the above conditions.*

When $\mathcal{GO}$ *is replaced by* $\mathcal{G}$ *in the above definition, one obtains the definition of an "*own-threat*". $\mathcal{A}_t$ will denote the set of all threats and own-threats that may be constructed from the bases* $<\mathcal{K}, \mathcal{G}, \mathcal{GO}>$.

With definition 5, the notion of own-threat covers both the own evaluation of P for the threats it receives, and the threats it may construct or imagine against itself from its own knowledge. Note that h may be a proposition whose truth can be controlled by the agent (e.g the result of an action), as well as a proposition which is out of its control. In a negotiation, an agent P may propose an offer x refused by O. In this case, the offer x is seen as an own-threat by O. P then entices O in order to accept the offer otherwise it will do an action which may be more painful for O. Here h is $Accept(x)$.

Definition 5 captures the two forms of threats. Indeed, in the first case (You should do 'a' otherwise I will do 'b'), $h =$ 'a', and in the second case (You should not do 'a' otherwise I will do 'b'), $h = \neg a$. 'b' refers to an action which may be inferred from H. The formal definition of threats is then slightly more general.

Example 3. *As said in example 1, the mother threatens her child not to let him go to the party organized by his friend if he does't finish his school work. The mother is supposed to have the following bases:*
$\mathcal{K}_{Mo} = \{\neg Work \rightarrow \neg Party\}$,
$\mathcal{G}_{Mo} = \{Work\}$,
$\mathcal{GO}_{Mo} = \{Party\}$.
The threat addressed by the mother to her child is formalized as follows: $<\{\neg Work \rightarrow \neg Party\}, Work, Party>$.

Let's now consider another dialogue between a boss and his employee.

Example 4

Boss: You should finish your work today.
Employee: No, I will finish it another day.
Boss: If you don't finish it you'll come this week-end to make overtime.

In this example, the boss has the three following bases:
$\mathcal{K}_{Bo} = \{\neg\ FinishWork \rightarrow Overtime\}$,
$\mathcal{G}_{Bo} = \{FinishWork\}$ *and*
$\mathcal{GO}_{Bo} = \{\neg Overtime\}$.
The threat enacted by the boss is: $<\ \{\neg\ FinishWork \rightarrow Overtime\}$, $FinishWork, \neg Overtime>$.

4.2 Strength of Threats

Threats involve goals and beliefs. Thus, the force of a threat depends on two criteria: the *certainty level* of the beliefs used in that threat, and the *importance* of the threatened goal.

Definition 6 (Force of a threat). *Let* $A = <H, h, \phi> \in \mathcal{A}_t$.
The force *of a threat A is a pair* $Force(A) = <\alpha, \beta>$ *s.t:* $\alpha = Level(H)$; $\beta = j$ *such that* $\phi \in \mathcal{GO}_j$.

However, when a threat is evaluated by its receiver (opponent), the threatened goal is in $\mathcal{G}$. In fact, the threatened goal may or may not be a goal of the opponent.

Definition 7 (Force of an own-threat). *Let* $A = <H, h, \phi> \in \mathcal{A}_t$.
The force *of an own-threat A is a pair* $<\alpha, \beta>$ *s.t.* $\alpha = Level(H)$; $\beta = j$ *if* $\phi \in \mathcal{G}_j$ *otherwise* $\beta = 0$.

Intuitively, a threat is strong if, according to the most certain beliefs, it invalidates an important goal. A threat is weaker if it involves beliefs with a low certainty, or if it only invalidates a goal with low importance. In other terms, the force of a threat represents to what extent the agent sending it (resp. receiving it) is certain that it will violate the most important goals of the other agent (resp. its own important goals). This suggests the use of a *conjunctive* combination of the certainty of H and the priority of the most important threatened goal. Indeed, a fully certain threat against a very low priority goal is not a very serious threat.

Definition 8 (Conjunctive combination). *Let A, B* $\in \mathcal{A}_t$ *with* $Force(A) = <\alpha, \beta>$ *and* $Force(B) = <\alpha', \beta'>$.
A is stronger *than B, denoted by* $A \succ_t B$, *iff* $min(\alpha, \beta) > min(\alpha', \beta')$.

Example 5. *Assume the following scale* $\{0, 1, 2, 3, 4, 5\}$. *Let us consider two threats A and B whose forces are respectively* $(\alpha, \beta) = (3, 2)$ *and* $(\alpha', \beta') = (1, 5)$. *In this case the threat A is* stronger *than B since* $min(3, 2) = 2$, *whereas* $min(1, 5) = 1$.

However, a simple conjunctive combination is open to discussion, since it gives an equal weight to the importance of the goal threatened and to the certainty of the set of beliefs that establishes that the threat takes place. Indeed, one may feel less threatened by a threat that is certain but has 'small' consequences, than by a threat which has a rather small plausibility, but which concerns a very important goal. This suggests to use a weighted minimum aggregation as follows:

Definition 9 (Weighted conjunctive combination). *Let A, B* $\in \mathcal{A}_t$ *with* $Force(A) = <\alpha, \beta>$, $Force(B) = <\alpha', \beta'>$.

A is stronger *than B,* $A \succ_t B$, *iff* $min(max(\lambda, \alpha), \beta) > min(max(\lambda, \alpha'), \beta')$, *where* λ *is the weight that discounts the certainty level component.*

The larger λ is, the smaller the role of α in the evaluation. The conjunctive combination is recovered when the value of λ is minimal.

Example 6. *Assume the following scale* $\{0, 1, 2, 3, 4, 5\}$. *Let us consider two threats A and B whose forces are respectively* $(\alpha, \beta) = (5, 2)$ *and* $(\alpha', \beta') = (2, 5)$. *Using a simple conjunctive combination, they both get the same evaluation* 2. *Taking* $\lambda = 3$, *we have* $min(max(3, 5), 2) = 2$ *and* $min(max(3, 2), 5) = 3$. *Thus B is stronger than A.*

The above approach assumes the commensurateness of three scales, namely the certainty scale, the importance scale, and the weighting scale. This requirement is questionable in principle. If this hypothesis is not made, one can still define a relation between threats.

Definition 10. *Let* $A, B \in \mathcal{A}_t$ *with* $Force(A) = <\alpha, \beta>$ *and* $Force(B) = <\alpha', \beta'>$.

A is stronger *than B iff:*

1. $\beta > \beta'$ *or,*
2. $\beta = \beta'$ *and* $\alpha > \alpha'$.

This definition also gives priority to the importance of the threatened goal, but is less discriminating than the previous one.

5 Rewards

During a negotiation an agent P can entice agent O in order that it does 'a' by offering to do an action 'b' as a reward. Of course, agent P believes that 'b'

will contribute to the goals of O. Thus, a reward has generally, at least from the point of view of its sender, a positive character. As for threats, two forms of rewards can be distinguished:

i) If you do 'a' then I will do 'b'.
ii) If you do not do 'a' then I will do 'b'.

The following example illustrates this idea.

Example 7. *A seller proposes to offer a set of blank CDs to a customer if this last accepts to buy a computer.*

5.1 Logical Definitions

Formally, rewards have an abductive form and are defined as follows:

Definition 11 (Reward). *A* reward *is a triple* $<H, h, \phi>$ *such that:*

1. h *is a propositional formula,*
2. $H \subseteq \mathcal{K}$,
3. $H \cup \{h\} \vdash \phi$ *such that* $\phi \in \mathcal{GO}$,
4. $H \cup \{h\}$ *is consistent and* H *is minimal (for set inclusion) among the sets satisfying the above conditions.*

When $\mathcal{GO}$ *is replaced by* $\mathcal{G}$ *in the above definition, one gets the definition of an* own-reward.

$\mathcal{A}_r$ *will denote the set of all the rewards that can be constructed from* $<\mathcal{K}, \mathcal{G}, \mathcal{GO}>$.

Note that the above definition captures the two forms of rewards. Indeed, in the first case (If you do 'a' then I will do 'b'), $h =$ 'a', and in the second case (If you do not do 'a' then I will do 'b'), $h = \neg a$.

Example 8. *Let's consider the example of a boss who promises one of his employee to increase his salary.*

Boss: You should finish this work ('a').
Employee: No I can't.
Boss: If you finish the work I promise to increase your salary ('b').

The boss has the following bases:
$\mathcal{K}_n = \{FinishWork \rightarrow IncreasedBenefit\}$,
$\mathcal{K}_{n-1} = \{IncreasedBenefit \rightarrow HigherSalary\}$,
$\mathcal{G}_n = \{FinishWork\}$ *and*
$\mathcal{GO}_n = \{HigherSalary\}$.
The boss presents the following reward in favor of its request 'Finish-Work': $<\{FinishWork \rightarrow HighBenefit,\ HighBenefit \rightarrow HighSalary\}, FinishWork, HighSalary>$.

Threats are sometimes thought as negative rewards. This is reflected by the parallel between the two definitions which basically differ in the third condition.

Remark 1. Let $\mathcal{K}$, $\mathcal{G}$, $\mathcal{GO}$ be the three bases of agent P. If $h \in \mathcal{G} \cup \mathcal{GO}$, $<\emptyset, h, h>$ is both a reward and a threat.

The above property says that if h is a *common goal* of the two agents P and O, then $<\emptyset, h, h>$ can be both a reward and a threat, since the common goals jointly succeed or fail. This is either both a reward and a own-reward, or a threat or a own-threat for P.

5.2 Strength of Rewards

As for threats, rewards involve beliefs and goals. Thus, the force of a reward depends also on two criteria: the certainty level of its support and the importance of the rewarded goal.

Definition 12 (Force of a reward). *Let $A = <H, h, \phi> \in \mathcal{A}_r$.*
The force *of a reward A is a pair $Force(A) = <\alpha, \beta>$ s.t: α = Level(H); $\beta = j$ such that $\phi \in \mathcal{GO}_j$.*

However, when a reward is evaluated by its receiver (opponent), the rewarded goal is in $\mathcal{G}$. In fact, if the proponent does not misrepresent the opponent's goals, the rewarded goal is a goal of the opponent.

Definition 13 (Force of an own-reward). *Let $A = <H, h, \phi> \in \mathcal{A}_t$. The* force *of an own-reward A is a pair $<\alpha, \beta>$ s.t. α = Level(H); $\beta = j$ if $\phi \in \mathcal{G}_j$, otherwise $\beta = 0$.*

Example 9. *In example 8, the force of the reward*
$<\{FinishWork \rightarrow HighBenefit, HighBenefit \rightarrow$
$HighSalary\}$, FinishWork, HighSalary> is <n-1, n>.

A reward is strong when for sure it will contribute to the achievement of an important goal. It is weak if it is not sure that it will help to the achievement of an important goal, or if it is certain that it will only enable the achievement of a non very important goal. Formally:

Definition 14 (Conjunctive combination). *Let A, B be two rewards in $\mathcal{A}_r$ with $Force(A) = <\alpha, \beta>$ and $Force(B) = <\alpha', \beta'>$.*
A is preferred *to B, denoted by $A \succ_r B$, iff $min(\alpha, \beta) > min(\alpha', \beta')$.*

However, as for threats, a simple 'min' combination is debatable, since it gives an equal weight to the importance of the rewarded goal and to the certainty of the set of beliefs that establishes that the reward takes place. Indeed, one may feel less rewarded by a reward that is certain but has 'small' consequences, than by a reward which has a rather small plausibility, but which concerns a very important goal. This suggests to use a weighted minimum aggregation as follows:

Definition 15 (Weighted conj. combination). *Let $A, B \in \mathcal{A}_r$ with $Force(A) =< \alpha, \beta >$ and $Force(B) =< \alpha', \beta' >$.*

$A \succ_r B$ iff $min(max(\lambda, \alpha), \beta) > min(max(\lambda, \alpha'), \beta')$, where λ is the weight that discounts the certainty level component.

The larger λ is, the smaller the role of α in the evaluation. The 'min' combination is recovered when the value of λ is minimal. In some situations, an agent may prefer a reward which is sure, even if the rewarded goal is not very important for it, than an uncertain reward with very 'valuable' consequences. This suggests to use a weighted minimum aggregation giving priority to the certainty component of the force, as follows:

Definition 16. *Let* $A, B \in \mathcal{A}_r$ *with* $Force(A) = <\alpha, \beta>$ *and* $Force(B) = <\alpha', \beta'>$.

$A \succ_r B$ *iff* $min(\alpha, max(\lambda, \beta)) > min(\alpha', max(\lambda, \beta'))$, *where* λ *is the weight that discounts the importance of the goal.*

Finally, as for threats, if there is no commensurateness of the three scales, we can still be able to compare two rewards as follows, in the spirit of definition 15:

Definition 17. *Let* $A, B \in \mathcal{A}_r$ *with* $Force(A) = <\alpha, \beta>$ *and* $Force(B) = < \alpha', \beta' >$.

$A \succ_r B$ *iff:*

1. $\beta > \beta$' *or,*
2. $\beta = \beta$' *and* $\alpha > \alpha$'.

This definition also gives priority to the importance of the rewarded goal. In the case of an agent which prefers rewards that are certain even if the rewarded goals are not very important, one can use the following preference relation.

Definition 18. *Let* $A, B \in \mathcal{A}_r$ *with* $Force(A) = <\alpha, \beta >$ *and* $Force(B) = < \alpha', \beta' >$.

$A \succ_r B$ *iff:*

1. $\alpha > \alpha$' *or,*
2. $\alpha = \alpha$' *and* $\beta > \beta$'.

6 Argumentation System

Due to the presence of potential inconsistency in knowledge bases, arguments may be conflicting. The most common conflict which may appear between explanatory arguments is the relation of *undercut* where the conclusion of an explanatory argument contradicts an element of the support of another explanatory argument. Formally:

Definition 19. *Let* $<H, h>, <H', h'> \in \mathcal{A}_e$. $<H, h>$ $defeats_e$ $<H', h'>$ *iff*

1. $<H, h>$ *undercuts* $<H', h'>$ *and*
2. *not* $(<H', h'> \succ_e <H, h>)$

Two threats may be conflicting for one of the three following reasons:

- the support of an argument infers the negation of the conclusion of the other argument. It occurs when, for example, an agent P threatens O to do 'b' if O refuses to do 'a', and at his turn, O threatens P to do 'c' if P does 'b'.

- the threats support contradictory conclusions. It occurs, for example, when two agents P and O have contradictory purposes.
- the threatened goals are contradictory. Since a rational agent should have consistent goals, $\mathcal{GO}$ should be as well consistent, and thus this arises when two threats are given by different agents.

As for threats, rewards may also be conflicting for one of the three following reasons:

- the support of an argument infers the negation of the conclusion of the other argument. It occurs when an agent P promises to O to do 'b' if O refuses to do 'a'. C, at his turn, promises to P to do 'c' if P does not pursue 'b'.
- the rewards support contradictory conclusions. This kind of conflict has no sense if the two rewards are constructed by the same agent. Because this means that the agent will contribute to the achievement of a goal of the other agent regardless what the value of h is. However, when the two rewards are given by different agents, this means that one of them wants h and the other $\neg h$ and each of them tries to persuade the other to change its mind by offering a reward.
- the rewarded goals are contradictory.

Formally:

Definition 20. *Let* $<H, h, \phi>, <H', h', \phi'> \in \mathcal{A}_t$ *(resp.* $\in \mathcal{A}_r$*).*
$<H', h', \phi'>$ $defeats_t$ $<H, h, \phi>$ *(resp.* $<H', h', \phi'>$ $defeats_r$ $<H, h, \phi>$*) iff*

1. $H' \vdash \neg h$*, or* $h \equiv \neg h'$*, or* $\phi \equiv \neg\phi'$*, and*
2. *not* $(<H, h, \phi> \succ_t <H', h', \phi'>)$ *(resp. not* $(<H, h, \phi> \succ_r <H', h', \phi'>)$*)*

It is obvious that explanatory arguments can conflict with threats and rewards. In fact, one can easily challenge an element used in the support of a threat or a reward. An explanatory argument can also conflict with a threat or a reward when the two arguments have contradictory conclusions. Lastly, an explanatory argument may conclude to the negation of the goal threatened (resp. rewarded) by the threat (resp. the reward). Formally:

Definition 21. *Let* $<H, h> \in \mathcal{A}_e$ *and* $<H', h', \phi> \in \mathcal{A}_t$ *(resp.* $\in \mathcal{A}_r$*).*
$<H, h>$ $defeats_m$ $<H', h', \phi>$ *iff*

1. $\exists h'' \in H'$ *such that* $h \equiv \neg h''$ *or*
2. $h \equiv \neg h'$ *or*
3. $h \equiv \neg\phi$.

Note that the force of the arguments is not taken into account when defining the relation "$defeat_m$". The reason is that firstly, the two arguments are of different nature. The force of explanatory arguments involves only beliefs while the the force of threats (resp. rewards) involves beliefs and goals. Secondly, beliefs have priority over goals since it is beliefs which determine whether a goal is justified and feasible.

Since we have defined the arguments and the conflicts which may exist between them, we are now ready to introduce the framework in which they are handled.

Definition 22 (Argumentation framework). *An* argumentation framework *is a tuple* $\langle \mathcal{A}_e, \mathcal{A}_t, \mathcal{A}_r, defeat_e, defeat_t, defeat_r, defeat_m \rangle$.

Any argument may have one of the three following status: *accepted*, *rejected*, or in *abeyance*. Accepted arguments can be seen as strong enough for having their conclusion, h, not challenged. In case of threats, for instance, an accepted threat should be taken seriously into account as well its logical consequences. Rejected arguments are the ones defeated by accepted one. Rejected threats will not be taken into account since they are too *weak* or not *credible*. The arguments which are neither accepted nor rejected are said in abeyance.

Let us define what is an accepted argument. Intuitively, accepted rewards (resp. threats) are the ones which are not defeated by another reward (resp. threat) or by an explanatory argument. Formally:

Definition 23 (Accepted threats/rewards). *Let* $\langle \mathcal{A}_e, \mathcal{A}_t, \mathcal{A}_r, defeat_e, defeat_t, defeat_r, defeat_m \rangle$ *be an argumentation framework.*

- *The* set of acceptable *threats is* $\mathcal{S}_t = \{A \in \mathcal{A}_t \mid \nexists B \in \mathcal{A}_t$ *(resp.* $\mathcal{A}_e$*),* $B\ defeats_t$ *(resp.* $defeats_m$*)* $A\}$. *A threat* $A \in \mathcal{A}_t$ *is* acceptable *iff* $A \in \mathcal{S}_t$.
- *The* set of acceptable *rewards is* $\mathcal{S}_r = \{A \in \mathcal{A}_r \mid \nexists\ B \in \mathcal{A}_r$ *(resp.* $\mathcal{A}_e$*),* $B\ defeats_r$ *(resp.* $defeats_m$*)* $A\}$. *A reward* $A \in \mathcal{A}_r$ *is* acceptable *iff* $A \in \mathcal{S}_r$.

7 Negotiation Protocol

As said in section 2, we suppose that we have two negotiating agents: P and O. Each of them has got a set $\mathcal{G}$ of *goals* to pursue, a knowledge base $\mathcal{K}$, and a base $\mathcal{GO}$, containing what the agent believes the goals of the other agent are. To capture the dialogues between these agents we follow [2] in using a variant of the dialogue system DC introduced by MacKenzie [8]. In this scheme, agents make dialogical moves by asserting facts into and retracting facts from *commitment stores* (CSs) which are visible to other agents. A commitment store CS is organized in two components: $CS.Off$ in which the *rejected offers* by the agent will be stored, and $CS.Arg$ which will contain the different arguments presented by the agent.

In addition to the different bases, each agent is supposed to be equipped with an argumentation system $\langle \mathcal{A}_e, \mathcal{A}_t, \mathcal{A}_r, defeat_e, defeat_t, defeat_r, defeat_m \rangle$. Note that the agent P constructs the arguments from the three following bases: $<\mathcal{K} \cup CS_C.Arg, \mathcal{G}, \mathcal{GO}>$.

The common agreement that negotiation aims to reach can be about a unique object or a concatenation of objects. Let X be the set of all possible offers. X is made of propositions or their negations.

7.1 Dialogue Moves

At each stage of the dialogue a participant has a set of legal moves it can make — making *offers*, accepting or rejecting offers, challenging an offer, presenting arguments, making threats or rewards. In sum, the set of allowed moves is

$\{Offer, Accept, Reject, Challenge, Argue, Threat, Reward\}$. For each move we describe how the move updates the CSs (the update rules), give the legal next steps possible by the other agent (the dialogue rules), and detail the way that the move integrates with the agent's use of argumentation (the rationality rules). In the following descriptions, we suppose that agent P addresses the move to the agent O.

Offer(x) where x is any formula in X. This allows the exchange of offers.

Rationality
- $\exists <H, x, \phi> \in S_r$ and it is an own-reward, and
- $<H, x, \phi> \succ_r <H', x', \phi'> \forall <H', x', \phi'> \in S_r$ and it is an own-reward with $x' \in X$.

In other terms, x is the most own-rewarding offer for the agent proposing it.

Dialogue: The other agent can respond with *Accept*(x), $Refuse(x)$, or $Challenge(x)$.

Update: There is no change.

$Challenge(x)$ where x is a formula in X.

Rationality: There is no rationality condition.

Dialogue: The other player can only $Argue(S, x)$ where $<S, x> \in \mathcal{A}_e$, or $Threat(H, x, \phi)$, or $Reward(H, x, \phi)$.

Update: There is no change.

After an offer, an agent can respond with

Accept(x) where $x \in X$.

Rationality: An agent P accepts an offer in one of the three following cases:
1. $\exists <H, x, \phi> \in S_r$ and it is an own-reward, and $<H, x, \phi> \succ_r <H', x', \phi'> \forall <H', x', \phi'> \in S_r$ and it is a own-reward with $x' \in X$, or
2. $\exists <H, x, \phi> \in \mathcal{S}_r$ and $<H, x, \phi> \in CS.Arg(O)$. This means that the agent has received an acceptable reward from the other agent.
3. $\exists <H, x, \phi> \in \mathcal{S}_t$ and $<H, x, \phi> \in CS.Arg(O)$. This means that the agent has been seriously threatened by the other agent.

Dialogue: The other player can make any move except $Refuse$.

Update: There is no change.

$Refuse(x)$ where x is any formula in X.

Rationality: $\exists <H, x, \phi> \in S_t$ and $<H, x, \phi>$ is a own-threat.

Dialogue: The other player can make any move except $Refuse$.

Update: $CS_i.Off(P) = CS_{i-1}.Off(P) \cup \{x\}$.

$Argue(A)$ where $A \in \mathcal{A}_e$, or $A \in \mathcal{A}_t$ or $A \in \mathcal{A}_r$.

Rationality: There is no rationality condition.

Dialogue: The other player can make any move except $refuse$.

Update: $CS_i.Arg(P) = CS_{i-1}.Arg(P) \cup \{A\}$.

$Threat(H, h, \phi)$ where $<H, h, \phi> \in \mathcal{A}_t$.

Rationality: $h \in CS.Off(O)$. This avoids that agents send gratuitous threats.
Dialogue: The other agent can respond with any move.
Update: $CS_i.Arg(P) = CS_{i-1}.Arg(P) \cup \{(H, h, \phi)\}$.

$Reward(H, h, \phi)$ where $<H, h, \phi> \in \mathcal{A}_r$.

Rationality: $h \in CS.Off(O)$. This avoids that agents send gratuitous rewards.
Dialogue: The other agent can respond with any move.
Update: $CS_i.Arg(P) = CS_{i-1}.Arg(P) \cup \{(H, h, \phi)\}$.

8 Illustrative Example

Let us illustrate the proposed framework in a negotiation dialogue between a boss B, and a worker W about finishing a work in time.

The knowledge base $\mathcal{K}_B$ of B is made of the following pieces of information, whose meaning is easy to guess ('overtime' is short for 'ask for overtime'):

$\mathcal{K}_n$ = {person-sick, overtime → finished-in-time, ¬ finished-in-time → penalty, finished-in-time → ¬ penalty, overtime-paid → extra-cost, strike → ¬ finished-in-time ∧ extra-cost}.
$\mathcal{K}_{a_1}$ = {person-sick → late-work},
$\mathcal{K}_{a_2}$ = {late-work ∧ ¬ overtime → ¬ finished-in-time}.

with $a_1 > a_2$. Goals of B are:

$\mathcal{G}_{b_1}$ = {¬ penalty},
$\mathcal{G}_{b_2}$ = {¬ extra-cost} with $b_1 > b_2$.

Moreover, for B,

$\mathcal{GO}_n$ = {overtime-paid},
$\mathcal{GO}_c$ = {¬ overtime}.

On his side, W has the following bases:

$\mathcal{K}_n$ = {overtime → late-work, overtime-paid → get-money},
$\mathcal{K}_{d_1}$ = {late-work ∧ overtime-paid → overtime},
$\mathcal{K}_{d_2}$ = {person-sick → late-work},
$\mathcal{K}_{d_3}$ = {¬late-work},
$\mathcal{K}_{d_4}$ = {¬ overtime-paid → strike},

with $d_1 > d_2 > d_3 > d_4$. Goals of W are

$\mathcal{G}_n$ = {overtime-paid},
$\mathcal{G}_{e_1}$ = {¬ overtime},

Finally, $\mathcal{GO}_f$ = {¬ strike}.

Possible actions (what is called the set of possible offers in the previous approach) for B are X = {overtime, ¬ overtime, overtime-paid, ¬overtime-paid}. Here it's a sketch of what can take place between B and W.

Step 1: B is led to make the move $Offer(overtime)$. Indeed, the agent can construct the following own-reward: <{overtime → finished-in-time, finished-in-time → ¬ penalty}, overtime, ¬penalty>. The force of this reward is $<n, b_1>$. Regarding $\neg overtime$, it can be checked that is not rewarding, and even threatening due to Th_1 = <{person-sick, person-sick→late-work, late-work ∧ ¬overtime →¬finished-in-time, ¬ finished-in-time →penalty}, ¬overtime, ¬penalty>, with the force $<\min(a_1, a_2), b_1>$. It can also be checked that *overtime* is most rewarding than the other actions in X.

Step 2: When W receives the command *overtime*, he makes the move $Challenge(overtime)$ because he can construct the own-threat $<\emptyset$, overtime, ¬overtime>. Moreover, the worker believes that he should't do overtime according to the explanatory argument <{overtime → late-work, ¬ late-work}, ¬overtime>.

Step 3: B makes the move Argue(Th_1) where he makes explicit to W his own-threat Th_1 used in step 1 for deciding his offer.

Step 4: Now W believes that there is effectively 'late-work' because he can construct the following accepted argument: <{person-sick, person-sick → late-work}, late-work>. Then he will suggest the offer 'overtime-paid' (Offer(overtime-paid)) because it is the most rewarding for him.

Step 5: B makes the move 'Refuse(overtime-paid)' since <{overtime-paid → extra-cost}, ¬overtime-paid, ¬extra-cost> is an own-threat for B.

Step 6: W threatens to go on strike. He presents the move $Threat(Th_2)$ with Th_2 = <{¬ overtime-paid → strike}, overtime-paid, ¬strike>.

Step 7: Th_2 is very serious by B. Indeed, two important goals of the agent will be violated if the worker executes that threat: $\neg penalty$ and ¬ extra-cost. In this case, B makes the move 'Accept(overtime-paid)' even if it is not acceptable for him.

9 Related Works – Conclusion

In [7], a list of the different kinds of arguments that may be exchanged during a negotiation has been addressed. Among those arguments, there are threats and rewards. The authors have then tried to define how those arguments are generated. They presented that in terms of speech acts having pre-conditions. Later on in [12], a way for evaluating the force of threats and rewards is given. However no formalization of the different arguments has been given, nor how their forces are evaluated, nor how they can be defeated.

In this paper we have presented a logical framework in which the arguments are defined. Moreover, the different conflicts which may exist between these arguments are described. Different criteria for defining the force of each kind of arguments are also proposed. Clearly, one may think of refining the criteria, especially by taking into account the number of threats or rewards induced by an offer, or the number of weak elements in the evaluation of certainty level. Since arguments may be conflicting we have studied their acceptability. We have also shown through a simple protocol how these arguments can be handled in a negotiation dialogue.

An extension of this work will be to study more deeply the notion of acceptability of such arguments. In this paper we have presented only the individual acceptability where only the direct defeaters are taken into account. However, we would like to investigate the notion of joint acceptability as defined in [5] in classical argumentation.

References

1. L. Amgoud and C. Cayrol. Inferring from inconsistency in preference-based argumentation frameworks. *Int. J. of Automated Reasoning*, 29:125–169, 2002.
2. L. Amgoud, N. Maudet, and S. Parsons. Modelling dialogues using argumentation. In *Proceedings of the International Conference on Multi-Agent Systems*, pages 31–38, Boston, MA, 2000.
3. L. Amgoud, S. Parsons, and N. Maudet. Arguments, dialogue, and negotiation. In *Proceedings of the 14th European Conference on Artificial Intelligence*, 2000.
4. L. Amgoud and H. Prade. Reaching agreement through argumentation: A possibilistic approach. In *9 th International Conference on the Principles of Knowledge Representation and Reasoning*, pages 194–201, Whistler, Canada, 2004.
5. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.
6. A. Kakas and P. Moraitis. Argumentative deliberation for autonomous agents. In *Proceedings of the ECAI'02 Workshop on Computational Models of Natural Argument (CMNA'02)*, pages 65–74, 2002.
7. S. Kraus, K. Sycara, and A. Evenchik. *Reaching agreements through argumentation: a logical model and implementation*, volume 104. Artificial Intelligence, 1998.
8. J. MacKenzie. Question-begging in non-cumulative systems. *Journal of philosophical logic*, 8:117–133, 1979.
9. S. Parsons, C. Sierra, and N. R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261—292, 1998.
10. I. Rahwan, S. D. Ramchurn, N. R. Jennings, P. McBurney, S. Parsons, and L. Sonenberg. Argumentation-based negotiation. *Knowledge engineering review*, 2004.
11. I. Rahwan, L. Sonenberg, and F. Dignum. Towards interest-based negotiation. In *AAMAS'2003*, 2003.
12. S. D. Ramchurn, N. Jennings, and C. Sierra. Persuasive negotiation for autonomous agents: a rhetorical approach. In *IJCAI Workshop on Computational Models of Natural Arguments*, 2003.