Threat, reward and explanatory arguments: generation and evaluation

Leila Amgoud[1] and **Henri Prade** [2]

Abstract. Current logic-based handling of arguments has mainly focused on explanation-oriented purposes in presence of inconsistency, so only one type of argument has been considered. Several argumentation frameworks have then been proposed for generating and evaluating such arguments. However, recent works on argumentation-based negotiation have emphasized different other types of arguments such as *threats*, *rewards*, *appeals*.
The purpose of this paper is to provide a logical setting which encompasses the classical argumentation-based framework and handles the new types of arguments. More precisely, we give the logical definitions of these arguments and 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 on the goals of other agents. In other words, from the different belief and goal bases maintained by agents, all the possible threats, rewards, explanations, appeals which are associated with them can be generated.
Key words: Negotiation, Argumentation

1 Introduction

Various argument-based frameworks have been developed in defeasible reasoning [1, 4, 7] for generating and evaluating arguments. In that explanation-oriented perspective, only one type of argument has been considered. Namely, what we call *explanatory* arguments. Recent works on negotiation [2, 5, 6, 8] have argued that argumentation plays a key role in finding a compromise. Indeed, an offer supported by a 'good argument' has a better chance to be accepted by another agent. Argumentation may also lead an agent to change its goals and finally may constrain an agent to respond in a particular way. For example, if an agent receives a threat, this agent may accept the offer even if it is not really acceptable for it. In addition to explanatory arguments studied in classical argumentation frameworks, the above works on argumentation-based negotiation have emphasized different other types of arguments such as *threats*, *rewards*, and *appeals*. In [5, 8], these arguments are treated as speech acts with pre-conditions and post-conditions.

The purpose of this paper is to provide a logical framework which encompasses the classical argumentation-based framework and handles the new types of arguments. More precisely, we give the logical definitions of these arguments and their weighting systems. These definitions take into account the fact that negotiation dialogues involve not only agents' beliefs (of various strengths), but also their goals (having maybe different priorities), and the beliefs on the goals of other agents. Thus, from the different belief and goal bases maintained by an agent, any possible threats, rewards, explanations, appeals, which are associated with them can be generated. Note that our weighting systems for threats and rewards reflect the certainty that they can take place and the importance of their consequences. However, they don't account for the propensity of the agent to act or not as it promise to do. An illustrative example is provided in the last section.

2 Types of arguments

In what follows, $\mathcal{L}$ denotes a propositional language, $\vdash$ classical inference, and $\equiv$ logical equivalence. In all what follows, we suppose also that an agent P presents an argument to another agent C. Each negotiating 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 finally a base $\mathcal{GO}$, containing what the agent believes the goals of the other agent are, as already assumed in [2]. $\mathcal{K}$ 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, levels of certainty are assigned to formulas in $\mathcal{K}$, and levels of priority are assigned to the goals. This leads to three possibilistic bases [3] that model gradual knowledge and preferences: $\mathcal{K} = \{(k_i, \alpha_i),\ i = 1, \ldots, n\}$, $\mathcal{G} = \{(g_j, \beta_j), j = 1, \ldots, m\}$, $\mathcal{GO} = \{(go_l, \delta_l), l = 1, \ldots, p\}$, where k_i, g_j, go_l are propositions of the language $\mathcal{L}$ and α_i, β_j, δ_l are elements of $[0, 1]$, or of any linearly ordered scale, finite or not. Both beliefs and goals are represented by propositional formulas. 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.
Each of these bases is associated at the semantics level with a possibility distribution which rank-orders the possible states of the world according either to their plausibility or to their satisfaction level. For instance, for the base $\mathcal{G}$ this distribution is a kind of qualitative utility function which is all the greater for a given state of the world as the corresponding interpretation does not violate any important goal in $\mathcal{G}$ (for more details on the semantics of possibilistic logic see [3]). We shall denote by $\mathcal{K}^*$, $\mathcal{G}^*$ and $\mathcal{GO}^*$ the corresponding sets of classical propositions when weights are ignored.
We distinguish between three categories of arguments according to their logical definitions: *threats*, *rewards* and *explanatory arguments*.

2.1 Threats

Threats are very common in human negotiation. They have a negative character and are applied to force an agent to behave in a certain way. Two forms of threats can be distinguished: 'You should do A otherwise I will do B' and 'You should not do A otherwise I will

[1] IRIT, 118 route de Narbonne, 31062 Toulouse Cedex, France
[2] IRIT, 118 route de Narbonne, 31062 Toulouse Cedex, France

CMNA IV, 4th *Workshop on Computational Models of Natural Argument*
Edited by Floriana Grasso, Chris Reed and Giuseppe Carenini
ECAI 2004

do B'. The first case occurs when an agent P needs an agent C to do A and C refuses P then threatens C to do B which, according to its beliefs, will have bad consequences for C. Let us consider the following example:

Example 1 *A mother asks her child (A) to carry out his school work and he refuses. The mother then threatens him (B) not to let him go to the party organized by her friend the next week-end.*

The second kind of threats occurs when an agent C wants to do some action A, which is not acceptable for P. In this case, P threatens that if C insists to do A then it will do B which, according to P's beliefs, will have bad consequences for C. To illustrate this kind of threat, we consider the following example borrowed from [5].

Example 2 *A labor union insists on having (A) a wage increase. The management says it cannot afford it, and asks the union to withdraw its request. The management threatens that, if it grants this increase, it will have (B) to lay off employees. According to the management, this will compensate for the higher operational cost that the increase will entail.*

In fact, 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. In the case of of example 1, the mother has a threat in favor of having the school work done. Formally:

Definition 1 (Threat) *A* threat *is a triple $<H, h, \phi>$ such that: 1) h is a proposition of the language $\mathcal{L}$, 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 minimal (for set inclusion) among the sets satisfying the conditions 1, 2, 3.*
$\mathcal{A}_t$ will denote the set of all threats that may be constructed from the bases $<\mathcal{K}, \mathcal{G}, \mathcal{GO}>$. H is the support *of the threat, h its* conclusion *and ϕ is the* threatened goal.

Such a definition allows h to 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 the control of the agent. For instance, "it rains and you are going to be wet". We may however restrict the set where h is taken, in order to exclude the last case. We may also allow for h = $\perp$ (the contradiction). This corresponds to the case of a gratuitous threat. Here actions are just literals whose conditional or unconditional effects are described through propositions stored in $\mathcal{K}$. Note that the above definition captures the two above 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.In example 1, H includes $\{\neg WorkDone \rightarrow \neg LettingGoParty\}$. The formal definition of threats is then slightly more general.

Example 3 *Let us consider an agent P having the three following bases: $\mathcal{K}$ = $\{(\neg\ FinishWork \rightarrow overtime, 1)\}$, $\mathcal{G}$ = $\{(FinishWork, 1)\}$ and $\mathcal{GO}$ = $\{(\neg overtime,\ 0.7)\}$. Let us suppose that the agent P asks the agent C to finish the work and that C refuses. P can then make the following threat: $<\{\neg\ FinishWork \rightarrow overtime\}, FinishWork, \neg overtime>$.*

2.2 Rewards

During a negotiation an agent P can entice agent C 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 C. 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: 'If you do A then I will do B' and 'If you do not do A then I will do B'.

Example 4 *A sales agent tries to persuade a customer (A) to buy a computer by offering (B) a set of blank CDs.*

Formally, a reward is defined in a way similar to threats as follows:

Definition 2 (Reward) *A* reward *is a triple $<H, h, \phi>$ such that: 1) h is a proposition of the language $\mathcal{L}$, 2) $H \subseteq \mathcal{K}^*$, 3) $H \cup \{h\} \vdash \phi$ such that $\phi \in \mathcal{GO}^*$, 4) $H \cup \{h\}$ is consistent and minimal (for set inclusion) among the sets satisfying the conditions 1, 2, 3. $\mathcal{A}_r$ will denote the set of all the rewards that can be constructed from $<\mathcal{K}$, $\mathcal{G}$, $\mathcal{GO}>$. H is the* support *of the reward, h its* conclusion *and ϕ the* rewarded goal.

Example 5 *Let's consider an agent P having the three following bases: $\mathcal{K}$ = $\{(FinishWork \rightarrow HighBudget,\ 1),\ (HighBudget \rightarrow HighSalary,\ 0.6)\}$, $\mathcal{G}$ = $\{(FinishWork, 1)\}$ and $\mathcal{GO}$ = $\{(HighSalary,\ 1)\}$. We suppose the agent P asks C to finish the work and C refuses. P can then present the following reward in favour of its request 'FinishWork': $<\{FinishWork \rightarrow HighBudget,\ HighBudget \rightarrow HighSalary\},\ FinishWork,\ HighSalary>$.*

In [5], another kind of arguments has been pointed out. It is the so-called *appeal to self-interest*. In this case, an agent P believes that the suggested offer implies one of C's goals. In fact, this case may be seen as a *self-reward* and consequently it is a particular case of rewards.

2.3 Explanatory arguments

Explanations constitute the most common category of arguments. In classical argumentation-based frameworks which have been developed for handling inconsistency in knowledge bases, each conclusion is justified by arguments. They represent the reasons to believe in the fact. Such arguments have a deductive form. Indeed, from premises, a fact or a goal is entailed. Formally:

Definition 3 (Explanatory argument) *An* explanatory *argument is a pair <H, h> such that: 1) $H \subseteq \mathcal{K}^* \cup \mathcal{G}^* \cup \mathcal{GO}^*$, 2) $H \vdash h$, 3) H is consistent and minimal (for set inclusion) among the sets satisfying the conditions 1 and 2. $\mathcal{A}_e$ will denote the set of all the explanatory arguments that can be constructed from $<\mathcal{K}$, $\mathcal{G}$, $\mathcal{GO}>$. H is the* support *of the argument and h its* conclusion.

Example 6 *Let us consider the case of an agent who wants to go to Sydney. $\mathcal{K}$ = $\{(conference,\ 0.8),\ (canceled,\ 0.4),\ (conference \rightarrow Sydney,\ 1),\ (canceled \rightarrow \neg conference,\ 1)\}$. $\mathcal{G}$ = $\{(Sydney, 1)\}$ and $\mathcal{GO}$ = $\emptyset$. The agent wants to go to Sydney and justifies his goal by the following explanatory argument: $<\{conference,\ conference \rightarrow Sydney\},\ Sydney>$. Indeed, from the beliefs one can deduce $Sydney$.*

In [5] other types of arguments called *appeals* are also considered. We argue that the different forms of appeals can be modeled as explanatory arguments. In what follows, we will show through examples how appeals can be defined in this way.

An appeal to prevailing practice: In this case, the agent believes that the opponent agent refuses to perform the requested action since it contradicts one of its own goals. However, the agent gives a counter-example from a third agent's actions, hoping it will serve as a convincing evidence. Of course, the third agent should have the same goals as the opponent and should have performed the action successfully.

Example 7 *An agent P asks another agent C to make overtime. C refuses because he is afraid that this is punished by law. The bases of C are then: $\mathcal{K}$ = $\{(overtime \rightarrow ToBePunished, 1)\}$, $\mathcal{G}$ = $\{(FinishWork, 1)\}$ and $\mathcal{GO}$ = $\emptyset$.*
When the opponent C receives the offer $overtime$, he constructs an explanatory argument in favor of $ToBePunished$: $<\{overtime, overtime \rightarrow ToBePunished\}, ToBePunished>$. This argument confirms to him that his goal will be violated and he refuses the offer. The proponent P reassures him by telling that another colleague makes overtime and he never has problems with the law. In fact, he presents the following counter-argument: $<\{overtime, \neg ToBePunished\}, \neg(overtime \rightarrow ToBePunished)>$. This last argument is an appeal to prevailing practice.

An appeal to past promise: In this case, the agent expects the opponent agent to perform an action based on past promise. Let us illustrate it by the following example:

Example 8 *A child asks his mother to buy a gift for him and the mother refuses. The child points out that she promised to buy something to him if he succeeds at his examinations. The bases of the child are: $\mathcal{K}$ = $\{(success, 1), (success \rightarrow \text{gift}, 1)\}$, $\mathcal{G}$ = $\{(\text{gift}, 1)\}$ and $\mathcal{GO}$ = $\emptyset$.*
The child's argument is then: $<\{success, success \rightarrow \text{gift}\}, \text{gift}>$.

A counter-example: This argument is similar to 'appeal to prevailing practice'; however, the counter-example is taken from the opponent agent's own history of activities. In this case, the counter argument produced by the proponent should be constructed from the beliefs of the opponent. In the case of example 7, the support of the counter-argument should be included in the base of C. Thus, C would have a conflicting base.

These three types of arguments have the same nature and they are all deductive. They are defined logically as explanatory arguments. The nature of these arguments, however, plays a key role in the strategies used by the agents. For example, a counter-example may quickly lead the other agent to change its mind than an appeal to prevailing practice.

As a conclusion of this section let us emphasize that a threat or a reward cannot be reduced to an explanatory argument as can be already seen on the definitions. On the one hand, explanatory arguments may lead the other agent to revise its beliefs / goals (they affect the mental states of the agent), while threats or rewards may encourage or refrain the agent to do something. On the other hand, the key entailment condition in the definition of threat, reward and explanatory arguments allows the following respective readings, H threatens ϕ provided $\neg h$, H rewards ϕ provided h and finally H explains h. Despite this apparent formal similarity, the two first expressions should be understood in a reverse way from an explanatory perspective. Indeed, in case of a threat or a reward this is rather the pair (H, ϕ) (although ϕ is the consequence of the entailment) which provides a kind of abductive explanation for h. Moreover, another important feature of definitions 1 and 2 is the requirement that ϕ belongs to $\mathcal{GO}^*$ which is distinct from $\mathcal{K}^*$ from which H is taken.

3 The strength of arguments

In [1], it has been argued that arguments may have different forces according to the beliefs from which they are constructed. The basic idea is that arguments using more certain beliefs are stronger than arguments using less certain beliefs. Thus, a level of certainty is assigned to each argument. These certainty levels make it possible to compare arguments. In fact, an argument A is preferred to another argument B iff A is stronger than B.
As mentioned before, each of the three bases $<\mathcal{K}, \mathcal{G}, \mathcal{GO}>$ is pervaded with uncertainty or equipped with priority levels. From these degrees, we first define the force of an explanatory argument.

Definition 4 (Force of an explanatory argument) *Let A = $<H, h> \in \mathcal{A}_e$. The* force *of $<H, h>$ is $Force(A)$ = $min\{a_i$ such that $(\varphi_i, a_i) \in H\}$.*

Example 9 *In example 6, the force of the explanatory argument $<\{conference, conference \rightarrow Sydney\}, Sydney>$ is 0.8. Whereas, the force of the argument $<\{canceled, canceled \rightarrow \neg conference\}, \neg conference>$ is 0.4.*

Concerning threats, things are different since a threat involves goals and beliefs. Intuitively, a threat is strong if, according to the most certain beliefs, it invalidates an important goal. A threat is weak if, according to less certain beliefs, it invalidates a less important goal. In other terms, the force of a threat represents to what extent the agent (the agent sending it or receiving it) is certain that it will violate its most important goals. Hence the use of a 'min' combination of the certainty of H and the priority of the threatened goal. Indeed, a fully certain threat against a very low priority goal is not a very serious goal. Formally:

Definition 5 (Force of a threat) *Let A = $<H, h, \phi> \in \mathcal{A}_t$. The* force *of a threat A is $Force(A)$ = $min(\alpha, \beta)$ such that α = $min\{a_i$ such that $(\varphi_i, a_i) \in H\}$ and $(\phi, \beta) \in \mathcal{GO} \cup \mathcal{G}$.*

Note that when a threat is evaluated by the proponent (the agent presenting the threat), then $(\phi, \alpha) \in \mathcal{G}O$. However, when it is evaluated by its receiver, $(\phi, \alpha) \in \mathcal{G}$.

Example 10 *In example 3 the force of the threat $<\{\neg FinishWork \rightarrow overtime\}, FinishWork, \neg overtime>$ is min(1, 0.7) = 0.7.*

As for threats, rewards involve beliefs and goals. Thus, a reward is strong when it is for sure that it will contribute to the achievement of an important goal. It is weak if it is not sure that it will contribute to the achievement of a less important goal.

Definition 6 (Force of a reward) *Let A = $<H, h, \phi> \in \mathcal{A}_r$. The* force *of a reward A is $Force(A)$ = $min(\alpha, \beta)$ such that α = $min\{a_i$ such that $(\varphi_i, a_i) \in H\}$ and $(\phi, \beta) \in \mathcal{GO} \cup \mathcal{G}$.*

Example 11 *In example 5, the force of the reward $<\{FinishWork \rightarrow HighBudget, HighBudget \rightarrow HighSalary\}, FinishWork, HighSalary>$ is 0.6.*

The forces of the arguments makes it possible to compare different arguments as follows:

Definition 7 (Preference relation) *Let A_1 and A_2 be two arguments of $\mathcal{A}_x$. A_1 is* preferred *to A_2, denoted by $A_1 \succ A_2$, iff $Force(A_1) \geq Force(A_2)$.*

In fact, the forces of arguments will play two roles: on the one hand they allow an agent to compare different threats or different rewards in order to select the "best" one. On the other hand, the forces are useful for determining the acceptable arguments among the conflicting ones.

4 Conflicts between arguments

Due to inconsistency in knowledge bases, arguments may be conflicting. In this section, we will show the different kinds of conflicts which may exist between arguments of the same nature and also between arguments of different natures.

In what follows, we denote by $\mathcal{A}_x$ the set of arguments of nature x with $x \in \{t, r, e\}$.

4.1 Conflicts between explanatory arguments

In classical argumentation frameworks, different conflict relations between what we call in this paper explanatory arguments have been defined. The most common ones are the relations of *rebuttal* where two explanatory arguments support contradictory conclusions and the relation of *undercut* where the conclusion of an explanatory argument contradicts an element of the support of another explanatory argument.

Definition 8 *Let $<H, h>$, $<H', h'> \in \mathcal{A}_e$. $<H, h>$ $defeats_e$ $<H', h'>$ iff $\exists$ h" $\in H'$ such that $h \equiv \neg h$", or $h \equiv \neg h'$.*

Example 12 (Continued) *In example 6, the explanatory argument $<\{canceled,\ canceled \rightarrow \neg conference\}, \neg conference>$ undercuts the argument $<\{conference,\ conference \rightarrow Sydney\}, Sydney>$ whereas it rebuts the argument $<\{conference\}, conference>$.*

4.2 Conflicts between threats / rewards

Two arguments of type "threat" 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. This case occurs when, for example, an agent P threatens C to do β if C refuses to do α, and at his turn, C threatens P to do δ if P does β.
- the threats support contradictory conclusions.
- the threatened goals are contradictory. Since a rational agent should have consistent goals, this case arises when the 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. This occurs when an agent P promises to C to do β if C refuses to do α. C, at his turn, promises to P to do δ if P does not pursue β.
- 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.

Definition 9 *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: $H' \vdash \neg h$, or $h \equiv \neg h'$, or $\phi \equiv \neg \phi'$.*

Note that the conflict relation between threats (or rewards) is generally symmetric.

4.3 Mixed conflicts

It is obvious that explanatory arguments can defeat threats and rewards. In fact, one can easily undercut an element used in the support of a threat or a reward. The defeat relation used in this case is the relation "undercut" defined above. An explanatory argument can also defeat a threat or a reward when the two arguments have contradictory conclusions. Finally, an explanatory argument may conclude the negation of the goal threatened (resp. rewarded) by the threat (resp. the reward). Formally:

Definition 10 *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: $\exists h$" $\in H'$ such that $h \equiv \neg h$" or $h \equiv \neg h'$ or $h \equiv \neg \phi$.*

5 Conclusion

The aim of this paper is twofold. First, it presents a logical framework in which the arguments are defined, the different conflicts which may exist between them are described, the force of each kind of arguments is defined in a clear way on the basis of the different bases of an agent and finally the acceptability of the arguments is studied. This work can be seen as a first formalization of different kinds of arguments. This is beneficial both for negotiation dialogue and also for argumentation theory since in classical argumentation the nature of arguments is not taken into account or the arguments are supposed to have the same nature.

REFERENCES

[1] L. Amgoud and C. Cayrol, 'A reasoning model based on the production of acceptable arguments', *Annals of Mathematics and Artificial Intelligence*, **34**, 197–216, (2002).

[2] 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, KR'2004*, (2004).

[3] D. Dubois, J. Lang, and H. Prade, 'Possibilistic logic', *Handbook of Logic in Artificial Intelligence and Logic Programming*, **3**, 439–513, (1993).

[4] 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).

[5] S. Kraus, K. Sycara, and A. Evenchik, *Reaching agreements through argumentation: a logical model and implementation*, volume 104, Journal of Artificial Intelligence, 1998.

[6] 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).

[7] H. Prakken and G. Sartor, 'Argument-based extended logic programming with defeasible priorties', *Journal of Applied Non-Classical Logics*, **7**, 25–75, (1997).

[8] 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).