Managing Deceitful Arguments with X-Logics

Geoffroy Aubry and **Vincent Risch**
InCA team, LSIS - UMR CNRS 6168
Domaine universitaire de Saint-Jérôme
13397 Marseilles cedex 20, France
email: {geoffroy.aubry, vincent.risch}@lsis.org

Abstract

In most works on negotiation dialogues, agents are supposed to be ideally honest. However, there are many situations where such a behaviour cannot always be expected from the agents (*e.g.* advertising, political negotiation, etc.). The aim of this paper is to reconsider the role of deceitful arguments in argumentation frameworks. We propose a logical tool for representing and handling deceitful arguments in a dialogue between two formal agents having to face their respective knowledge and trying to convince each other. X-logics, a nonmonotonic extension of classical propositional logics, is used as the background formalism for representing the reasoning of the agents on arguments. Starting from a previous work dedicated to the generation of new arguments, we propose to define the notion of lie as a new kind of possible agent's answer. Finally we describe the way an agent may trick and how the other agent may detect it.

Introduction

Logics has been historically assigned by philosophy the task of defining the rules of correct reasoning. As such, it was first considered as a tool in argumentation and rhetoric, before its formal developments in the foundation of mathematics meant to "delimit" it (in a large sense) to proof theory. A major distinction then arised from this situation: whereas the inner nature of arguments makes them questionable, theorems are held to be beyond dispute. A definitive break seemed to be set between logics and argumentation, leading a philosopher like Perelman to regard human argumentation as being beyond the reach of formal logics. New developments in logics however have conducted to reconsider partly this pessimistic view. In recent years the representation and the simulation of simplified models of argumentation with logical tools has been the object of important progresses (Prakken & Vreeswijk 2002). Obviously, all these models initially assume a form of exchange of arguments. In addition they address and characterize many distinct fundamental notions such as (for instance) acceptability (Dung 1995), preference among arguments (Amgoud & Cayrol 1998), argumentation trees (Besnard & Hunter 2001), relative strenght of arguments (Bench-Capon 2003), or dialectic proofs (Dung, Kowalski, & Toni 2006). Another important feature of these models is also that they assume some kind of rational behaviour in the exchange of arguments, that is that the agents (implicitly or explicitly) involved can be fully trusted. However, there are many contexts in which this assumption seems much too strong, typically situations in which the object of the discussion covers important issues for at least one part (*e.g.* financial or political negotiation, or yet, advertising). Actually the purpose of argumentation as a part of rhetoric is indeed to propose and stress values to which each agent believes or *feint to believe* regarding her respective goals. In this respect, the ability to represent, manage, and detect deceitful arguments appears as a major step toward a complete formal theory of argumentation, and as such, was already questioned in (Hamblin 1970)'s pioneering work. In this paper, and following (Aubry & Risch 2005), we address this question via the use of X-logics, a nonmonotonic extension of classical propositional logics. Extending their approach, devoted to the question of the generation of new arguments, we propose a logical approach to the notion of deceitful argument, and show how agents may generate lies while maintaining the consistency of their knowledge. The question of the detection of lies is briefly considered via the notion of *commitment store* (Hamblin 1970). Our paper is organized as follows: section 2 below briefly introduces standard notations, section 3 recalls X-logics, section 4 introduces the notions of agent, attitudes, kinds of answers and arguments, while section 5 concerns deceitful arguments.

Notations

Formally our language is classical propositional logic denoted by $\mathcal{L}$. Formulas are denoted by lowercase letters whereas sets of formulas are denoted by shift case letters. The symbols $\top$ and $\bot$ are the usual truth values, and $\neg$, $\vee$, $\wedge$, $\Rightarrow$, $\Leftrightarrow$ the usual connectors. Classical consequence relation is denoted by $\vdash$. A finite set E of formulas is logically interpreted by the conjunction of its elements, that is a sentence. We abuse the notation $\neg E$ as a shorthand for the negation of the conjunction of the formulas in E, *e.g.* $E = \{e_1, \ldots, e_n\}$, hence $\neg E = \neg e_1 \vee \cdots \vee \neg e_n$. We denote by $\overline{E}$ the set of classical consequences of E (*i.e.* $\overline{E} = \{f \mid E \vdash f\}$), and by 2^E, the powerset of E. A finite *consistent* set of formulas is called a *knowledge base*.

X-logics

X-logics were defined in (Siegel & Forget 1996) as an attempt for defining a proof theory for nonmonotonic logics from any classical logic with a given set X of formulas. Whereas classically $K \vdash f$ iff $\overline{K \cup \{f\}} = \overline{K}$, X-logics can be considered as a generalization (hence a weakening) of $\vdash$, namely $\vdash_X$, defined such that $K \vdash_X f$ iff $\overline{K \cup \{f\}} \cap X = \overline{K} \cap X$, *i.e.* $\vdash_X$ is monotonic only on X. When $X = \mathcal{L}$, $\vdash_X$ amounts to be just $\vdash$. If $X = \{\bot\}$ then $K \vdash_X f$ is equivalent to $K \nvdash \neg f$ which describes the consistency relation between K and f ("$K \wedge f$ is satisfiable" holds), provided K is consistent by itself. If $X = \varnothing$, all the formulas can be entailed. Note that $K \vdash_X f$ if every theorem (regarding $\vdash$) of $K \cup \{f\}$ which is in X is a theorem of K (by adding f to K the set of classical theorems which are in X does not grow). Indeed, since classical consequence relation is monotonic, in order to check whether $K \vdash_X f$ it is sufficient to check whether $\overline{K \cup \{f\}} \cap X \subseteq \overline{K}$. In other words, $K \vdash_X f$ iff $\forall x \in X \setminus \overline{K}$, $K \cup \{f\} \nvdash x$. Although this was already proved independently, this shows that X-logics are supraclassical. Actually and as shown in (Bochman 2003), X-logics coincide with permissive inference relations which are completely characterized by Left Logical Equivalence, Right Weakening, Reflexivity, Conjunctive Cautious Monotony, Cut and Or.

Let us make use of the following terminology: if $K \vdash_X f$ we say that f is *compatible* with K regarding X, and *incompatible* otherwise. The notion of compatibility encompasses the notion of consistency, whereas formulas can be incompatible with K regarding X without being inconsistent with K. The following properties obviously hold:

Property 1.

1. (metacoherence) *A formula cannot be both compatible and incompatible.*
2. (paraconsistency of compatibility) *Both a formula and its negation can be compatible with K regarding X.*
3. (paraconsistency of incompatibility) *Both a formula and its negation can be incompatible with K regarding X.*

Example 2.

- $\{a\} \vdash_{\{\bot\}} a \wedge b$, and $\{a\} \vdash_{\{\bot\}} \neg(a \wedge b)$
- $\{a\} \nvdash_{\{\bot, b, \neg b\}} a \wedge b$, and $\{a\} \nvdash_{\{\bot, b, \neg b\}} \neg(a \wedge b)$
- $\{a\} \vdash_{\{b \wedge c\}} b$, but $\{a, c\} \nvdash_{\{b \wedge c\}} b$

Agents, attitudes, answers and arguments

In the literature, some argumentation theories consider the notion of *proponent-opponent* (Rescher 1977; Vreeswijk 1992) whereas other describe argumentation systems in which arguments made from a unique set of formulas are linked together, in a kind of abstract game among arguments (Lin & Shoham 1989; Dung 1995; Amgoud & Cayrol 1998; Besnard & Hunter 2001). Following (Simari & Loui 1992; Amgoud & Parsons 2002), (Aubry & Risch 2005), we introduce a notion of *agent*, but with the objective to map each agent with a unique X-inference.

Definition 3. *(Aubry & Risch 2005) An* agent *is a couple $[K, X]$ where K is a knowledge base, and $X \supseteq \{\bot\}$, a set of formulas. The set of agents, a subset of $2^{\mathcal{L}} \times 2^{\mathcal{L}}$, is denoted by $\mathcal{A}$.*

Compatibility extends naturally to the notion of admissibility of a formula by an agent, *i.e.* a formula is *admissible* by an agent $[K, X]$ iff this formula is compatible with K regarding X; it is *non-admissible* otherwise [1]. Intuitively K is used as a representation of the factual knowledge of an agent, whereas X corresponds to formulas that an agent cannot admit unless they are part of her factual knowledge. In other words, formulas in X delineate negatively the hopes of the agent (since the agent does not admit these formulas), whereas the positive counterpart indeed should correspond to the agent's expectations (the agent accept everything but formulas of X, unless she has to take account of them because there are already part of her knowledge). In the following, we will call X the *forbidden formulas*. The requirement that X contains at least the contradiction is motivated by the natural expectation that an agent should reason consistently. The notion of admissibility determines four possible distinct *attitudes* that an agent may adopt concerning a given formula, as shown in (Aubry & Risch 2005):

Definition 4 (Attitudes)**.** *(Aubry & Risch 2005) Consider an agent $[K, X]$ and a formula f:*

- $[K, X]$ *is* for f *iff* $K \vdash_X f$ *and* $K \nvdash_X \neg f$
- $[K, X]$ *is* neutral *about* f *iff* $K \vdash_X f$ *and* $K \vdash_X \neg f$
- $[K, X]$ *is* puzzled *by* f *iff* $K \nvdash_X f$ *and* $K \nvdash_X \neg f$
- $[K, X]$ *is* against f *iff* $K \nvdash_X f$ *and* $K \vdash_X \neg f$

By extension, an agent is for *(resp.* neutral *about,* against, puzzled *by) a set of formulas iff she is* for *(resp.* neutral *about,* against, puzzled *by) the conjunction of the formulas of this set.*

In figure 1, the four corners of the median layout are associated with the attitudes, and are clearly generated by both the two edges above and below corresponding to the admissible or non-admissible character of f and $\neg f$ respectively.

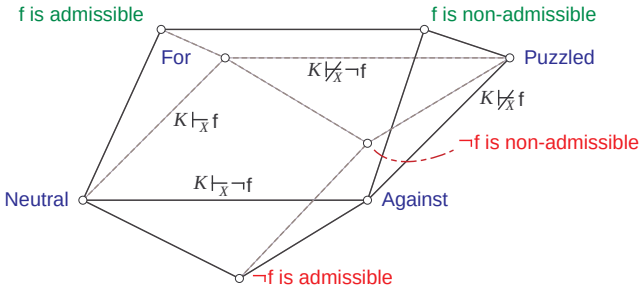

Figure 1: Generative octahedron of attitudes of an agent $[K, X]$ in front of a formula f

[1] Note that our notion of admissibility differs from the notion previously defined in (Bondarenko *et al.* 1997) and in (Dung, Kowalski, & Toni 2006).

Yet, as shown in (Aubry & Risch 2005), given an agent Φ and a formula f:

- Φ is for f iff she is against $\neg f$,
- Φ is neutral about f iff she is neutral about $\neg f$,
- Φ is puzzled by f iff she is puzzled by $\neg f$,
- Φ is for the tautologies and against the contradictions.

The different possible attitudes of an agent in front of a set of formulas yield a partition of this set. *Confrontation operators* are introduced in order to associate each attitude with the formulas of a particular partition.

Definition 5. *(Aubry & Risch 2005) The* operator $|_+$ *(resp.* $|_0$, $|_-$ *and* $|_p$*) maps an agent and a set* E *of formulas with the subsets of* E *such that this agent is* for *(resp.* neutral *about,* against *or* puzzled *by) these subsets:*

$$\begin{array}{rccl} |_+ : & \mathcal{A} \times 2^{\mathcal{L}} & \longrightarrow & 2^{2^{\mathcal{L}}} \\ & \Phi \mid_+ E & \longmapsto & \{P \subseteq E \mid \Phi \textit{ is for } P\} \end{array}$$

As shown further down, confrontation operators are meant to be used by an agent for deciding precisely which are the points of agreement or disagreement she has with the different parts of a given argument.

Example 6. Consider Φ an agent with the following knowledge base and set of forbidden formulas (where A and B denote respectively "Amelia" and "Brandon"): $K_\Phi = \{$*B-comes* $\Rightarrow$ *Annoyed*, $\neg$*B-comes* $\Rightarrow$ *Sad*, *A-comes* $\Leftrightarrow$ *Happy*$\}$, $X_\Phi = \{\bot,$ *Annoyed*, *Sad*$\}$. The attitudes of Φ regarding the set $E = \{$*B-comes*, *A-comes*, *B-comes* $\Rightarrow$ *Annoyed*, *A-comes* $\wedge$ $\neg$*Happy*$\}$ yield the following partition of E:

$\Phi \mid_+ E = \big\{\{$*B-comes* $\Rightarrow$ *Annoyed*$\}\big\}$
$\Phi \mid_0 E = \big\{\{$*A-comes*$\}, \{$*A-comes*, *B-comes* $\Rightarrow$ *Annoyed*$\}\big\}$
$\Phi \mid_- E = \big\{\{$*A-comes* $\wedge$ $\neg$*Happy*$\}, \{$*B-comes*, *A-comes*$\}, *\big\}$
$\Phi \mid_p E = \big\{\{$*B-comes*$\}, \{$*B-comes*, *B-comes* $\Rightarrow$ *Annoyed*$\}\big\}$

The symbol $*$ in $\Phi \mid_- E$ stands for each subset of E containing either of the two other sets stated in $\Phi \mid_- E$.

In order to link the attitudes of an agent with the construction of new arguments (Aubry & Risch 2005) make use of the following notion of *answer*. Roughly, an answer is a set of formulas fixed by the attitude of an agent regarding a given set of formulas this agent is faced with. Consider $X = \{x_1, \ldots, x_n\}$, and let us use conceivable (X) as a shorthand for $\{\neg x_1, \ldots, \neg x_n\}$.

Definition 7 (Answer)**.** *(Aubry & Risch 2005) An* answer *of the agent* $[K, X]$ *to a consistent set* A *of formulas is a consistent set* R *of formulas such that, for some* $K' \subseteq K$ *and some* $X' \subseteq (X \setminus \{\bot\})$*:*

1. $R = K' \cup$ conceivable (X')
2. $K' \nvdash_{\{\bot\} \cup X'} A$

The set of answers given by $[K, X]$ *to* A *is written* $\mathcal{R}^A_{[K,X]}$.

The first point constrains answers to contain only knowledge or negations of forbidden formulas of the agent. The second point specifies that an answer to a set A necessarily contains formulas which are conflicting with A: $\exists x \in (\{\bot\} \cup X') \setminus \overline{K'},\ K' \cup A \vdash x$. In other words, an answer $K' \cup$ conceivable (X') to A is such that A is non-admissible by the "virtual" agent $[K', \{\bot\} \cup X']$. The contradiction is present here for technical reasons: for instance, this allows to answer $R = K' = \{a\}$ to $A = \{\neg a\}$ without having to add $X' = \{\bot\}$ in R.

Following directly from definition 7: *(1)* no answer is empty, *(2)* any answer of an agent to a given set of formulas is inconsistent with this set. In addition, the following property holds:

Property 8 (Existence of an answer)**.** *(Aubry & Risch 2005) If an agent is* against *or* puzzled *about a subset of a consistent set of formulas, then there exists an answer of this agent to this set. The opposite does not hold.*

In the following, we consider two easy but important refinements of definition 7, namely *coherent*, and *relevant* answers. Let us define the first of these two notions:

Definition 9 (Coherent answer)**.** *An answer of the agent* $\Phi = [K, X]$ *is called a* coherent answer *of* Φ *iff it is consistent with* K.

As stated further down, each answer of an agent Φ to A (and especially coherent answers) is potentially the support of a counterargument of Φ (that is the reason why to believe the conclusion of this counterargument) to an argument containing A. Note however this does not limit agents to generate only disputing arguments (remind that Φ is for A iff she is against $\neg A$). Answers that are not coherent will drive us to deceitful arguments, and as such definition 9 plays an important role as the counterpart of the notion of lie, considered further down.

Let us now come to the second refinement of definition 7. In most works on argumentation, only minimal arguments are considered: they only contains formulas necessary to elaborate the conclusion of the argument. Among the possible answers of an agent to a set of formulas A, some of them are included in others: they contain less superfluous information. In the limit case, the minimal answers are only made with formulas necessary to elaborate the conclusion $\neg A$. Such answers are called *relevant* [2].

Definition 10 (Relevant answer)**.** *(Aubry & Risch 2005) An answer of the agent* Φ *to a set* A *of formulas is called* relevant *iff it does not contain any other answer of* $\mathcal{R}^A_\Phi$*. The set of relevant answers given by* Φ *to* A *is written* $\mathcal{R}r^A_\Phi$.

$$\mathcal{R}r^A_\Phi \ = \ \big\{R \in \mathcal{R}^A_\Phi \mid \forall R' \subset R,\ R' \notin \mathcal{R}^A_\Phi\big\}$$

Now, given an agent $[K, X]$ facing a set A of formulas, her set of relevant answers can be shared among three subsets. The first contains the answers only made from the knowledge of this agent. Hence this subset is not empty when $K \cup A$ is inconsistent. The second category of answers is made of those exclusively constructed from the forbidden

[2] Note that the notion of *relevant move* previously defined in (Prakken 2005) has a different meaning, since it is defined in the context of a dialogue.

formulas of this agent. Finally, the relevant answers neither in the first nor in the second subset are made from at least one formula of the set of knowledge of this agent and one formula of the set of forbidden formulas of this agent.

This partitioning can be considered via "virtual" agents. The answers only made from the knowledge of the agent $[K, X]$ are actually answers of an agent $[K, \{\bot\}]$, while the answers only made of the forbidden formulas of the agent $[K, X]$ are answers of the agent $[\varnothing, X]$.

These different sets of relevant answers are fully characterized in (Aubry & Risch 2005).

Answers are used as a tool for generating new arguments by an agent, where arguments are defined following a very common intuitive view (Simari & Loui 1992; Elvang-Gøransson, Krause, & Fox 1993; Amgoud & Cayrol 1998; Besnard & Hunter 2001), *i.e.* an *argument* is a set of relevant formulas that can be used to classically prove some formula, together with that formula. This notion is made more precise here by taking account of both notions of agent and answer.

Definition 11 (Argument)**.** *(Aubry & Risch 2005) Consider A, a set of formulas. An* argument *α of an agent Φ is any pair $\langle R, \neg A\rangle$ such that R is a relevant answer of Φ to A. The set of arguments of an agent Φ is written $\mathcal{A}rg_\Phi$, i.e.*

$$\forall \Phi, \forall R, \forall A, \quad \langle R, \neg A\rangle \in \mathcal{A}rg_\Phi \text{ iff } R \in \mathcal{R}r_\Phi^A$$

The set of all arguments is denoted by $\mathcal{A}rg$. Finally, R is called the support *of the argument, denoted by* $\mathsf{supp}(\alpha)$*, while $\neg A$ is called the* conclusion *of the argument, denoted by* $\mathsf{concl}(\alpha)$.

From the definition of an answer (definition 7), we have indeed that the conclusion of an argument is classically entailed by the support of this argument. Let us now address the question of the use of arguments by an agent. The two complementary notions of *attack* and *defense* of an argument in the context of a dispute are well known in philosophy. Following a solid tradition (*e.g.* (Schopenhauer 2004)) we consider that an argument can be attacked (resp. defended) either on the premises (the support) or on the conclusion. Hence we allow arguments to be decomposed into *elements* that an agent can analyze for such further attack or defense. Thus, the elements of an argument are taken as parts of the support, together with the conclusion:

Definition 12 (Elements of an argument)**.** *The* elements *of an argument $\langle S, c\rangle$, written* $\mathsf{elements}(\langle S, c\rangle)$*, are given by a mapping from $\mathcal{A}rg$ to $2^{2^\mathcal{L}}$ such that:*

$$\mathsf{elements}(\langle S, c\rangle) = \{E \in 2^\mathcal{L} \mid E \subseteq S\} \cup \{\{c\}\}$$

The two relations of attack and defense are then defined classically. Note however that generally, defending arguments is considered through *reinstatement*: an argument that is defeated by another argument can be justified only if it is reinstated by a third argument (this corresponds to (Dung 1995)'s notion of *acceptability*). Actually, the notions of attitude and the generation of agent's answers allow us to define a pure relation of defense of arguments.

Definition 13. *The set of arguments of an agent Φ* attacking *(resp.* defending*) an argument α is written $\mathcal{A}rg_\Phi^{\mathsf{att}(\alpha)}$ (resp. $\mathcal{A}rg_\Phi^{\mathsf{def}(\alpha)}$), where:*

- $\mathcal{A}rg_\Phi^{\mathsf{att}(\alpha)} = \{\langle R, \neg A\rangle \mid R \in \mathcal{R}r_\Phi^A,\ A \in \mathsf{elements}(\alpha)\}$
- $\mathcal{A}rg_\Phi^{\mathsf{def}(\alpha)} = \{\langle R, \bigwedge_{a\in A} A\rangle \mid R \in \mathcal{R}r_\Phi^{\{\neg A\}}, A \in \mathsf{elements}(\alpha)\}$

The following property links the attack and the defense of arguments with the attitudes of an agent:

Property 14. *Consider $t \in \{\mathsf{supp}(), \mathsf{concl}()\}$. $\forall \Phi, \forall A, \forall \alpha$:*

- $A \in \Phi \mid_- t(\alpha) \Rightarrow \exists R,\ \langle R, \neg A\rangle \in \mathcal{A}rg_\Phi^{\mathsf{att}(\alpha)}$
- $A \in \Phi \mid_+ t(\alpha) \Rightarrow \exists R,\ \langle R, \bigwedge_{a\in A} a\rangle \in \mathcal{A}rg_\Phi^{\mathsf{def}(\alpha)}$
- $A \in \Phi \mid_p t(\alpha) \Rightarrow \begin{cases} \exists R_1, \exists R_2, \\ \langle R_1, \neg A\rangle \in \mathcal{A}rg_\Phi^{\mathsf{att}(\alpha)} \\ \langle R_2, \bigwedge_{a\in A} a\rangle \in \mathcal{A}rg_\Phi^{\mathsf{def}(\alpha)} \end{cases}$

Sketch of proof. Applying property 8 then definition 11. □

Note that property 14 makes our intuition about the attitudes of an agent to coincide with her expected behaviour: an agent is puzzled by one element of an argument because she can both attack and defend this argument. Similarly, the fact that an agent is neutral about some element of an argument does not allow this agent to construct any coherent answer, and hence does not allow her to attack or to defend this argument (unless by lying as seen further down).

Example 15. Consider two agents Φ and Ψ arguing about the text of the future European constitution (denoted by x) in order to decide whether x should be accepted or not. The knowledge of Φ regarding x is that x mentions the concept of trade market, that x pretends to be a constitution as well as to set up new political foundations for Europe, and that the need of political foundations requires a constitution anyway. On the other hand, Φ is not ready to give up the idea that a constitution should not include any reference to trade markets. Hence, we have $K_\Phi = \{x, x \Rightarrow \textit{tradeM}, x \Rightarrow \textit{constitution}, x \Rightarrow \textit{newPoliticalF}, \textit{newPoliticalF} \Rightarrow \textit{constitution}\}$, and $X_\Phi = \{\bot, \neg((\textit{constitution} \wedge \textit{tradeM}) \Rightarrow \neg\textit{admissible})\}$. The knowledge of the agent Ψ is that x refers to trade markets, and as such, has to be considered a treaty rather than a constitution. Moreover Ψ thinks that a treaty can be accepted. Hence, $K_\Psi = \{x, x \Rightarrow \textit{tradeM}, x \Rightarrow \textit{treaty}, \textit{treaty} \Leftrightarrow \neg\textit{constitution}, \textit{treaty} \Rightarrow \textit{admissible}\}$, while $X_\Psi = \{\bot\}$.

Assume now the claim made by Ψ in front of Φ that x should be accepted, *i.e.* $\alpha_\Psi^1 = \langle\{x \Rightarrow \textit{treaty}, \textit{treaty} \Rightarrow \textit{admissible}\}, x \Rightarrow \textit{admissible}\rangle$.

Since $\{x \Rightarrow \textit{admissible}\} \in \Phi \mid_- \mathsf{concl}(\alpha_\Psi^1)$, Φ can compute a counterargument: $\alpha_\Phi^1 = \langle\{x, x \Rightarrow \textit{tradeM}, x \Rightarrow \textit{constitution}, (\textit{constitution} \wedge \textit{tradeM}) \Rightarrow \neg\textit{admissible})\}, \neg(x \Rightarrow \textit{admissible})\rangle$.

Now Ψ analyzes α_Φ^1: $\Psi \mid_+ \mathsf{supp}(\alpha_\Phi^1) = \{\{x\}, \{x \Rightarrow \textit{tradeM}\}, \{(\textit{constitution} \wedge \textit{tradeM}) \Rightarrow \neg\textit{admissible})\}, *\}$, where $*$ stands for each subset of the set containing $*$, and Ψ

is against any other elements of α^1_Φ. Ψ set up the following argument: $\alpha^2_\Psi = \langle\{x, x \Rightarrow \textit{treaty}, \textit{treaty} \Leftrightarrow \neg\textit{constitution}\}, \neg(x \Rightarrow \textit{constitution})\rangle$. On her turn, Φ defends her position: $\alpha^2_\Phi = \langle\{x \Rightarrow \textit{newPoliticalF}, \textit{newPoliticalF} \Rightarrow \textit{constitution}\}, \neg\neg(x \Rightarrow \textit{constitution})\rangle$.

Deceitful arguments

Deceitful arguments are untruthful arguments made with the intention to deceive. An agent using such kind of argument cheats with her current knowledge about the world, and does so with the intention to trick the other agent. We consider here two ways of cheating: the first one has to do with some kind of dilution of an argument, whereas the second one refers directly to lies. Let us informaly consider dilution first.

Given an initial argument, in order to narrow the possibilities of counterarguments to consider, (Besnard & Hunter 2001) introduce the following notion of *conservativity*:

Definition 16. *(Besnard & Hunter 2001) An argument $\langle S, c\rangle$ is* more conservative *than an argument $\langle S', c'\rangle$ iff $S \subseteq S'$ and $c' \vdash c$.*

Selecting only the most conservative counterarguments allows to summarize the different possibilities of answer (as shown in (Aubry & Risch 2005)). However, depending on the goal of the agent beginning a discussion, the most conservative argument may not be the most suitable one as first argument:

Example 17. Consider a commercial agent Φ who wants sell a scooter of trademark *label-Z* to a client Ψ. $\Phi = [\{\textit{scooter}, \textit{label-Z}, \textit{scooter} \Rightarrow \textit{edgeOut}, \textit{edgeOut} \Rightarrow \neg\textit{trafficJam}\}, \{\bot\}]$.

Φ uses the following argument: $\langle\{\textit{scooter} \Rightarrow \textit{edgeOut}, \textit{edgeOut} \Rightarrow \neg\textit{trafficJam}\}, (\textit{scooter} \wedge \textit{label-Z}) \Rightarrow \neg\textit{trafficJam}\rangle$. But the agent Ψ considers it as a *diluted argument* since she does not understand why an unsophisticated scooter is not enough to avoid the traffic jams. She can then address the following argument to the commercial agent: $\langle\{\textit{scooter} \Rightarrow \textit{edgeOut}, \textit{edgeOut} \Rightarrow \neg\textit{trafficJam}\}, \textit{scooter} \Rightarrow \neg\textit{trafficJam}\rangle$. Since Φ is for each element of this argument, she cannot generate any counterargument.

Hence depending on how much Ψ is careful, the first argument of Φ seems to be useless while a less conservative argument may appear more appropriate. A new kind of answer can assist our commercial agent to stick up for herself, which drive us two the second way by which an agent may cheat:

Definition 18 (Lie)**.** *M is a* lie *of the agent Φ regarding the set of formulas A iff both M is a relevant answer of Φ to A, and M is not a coherent answer of Φ. The set of lies of Φ regarding A is denoted by $\mathcal{RL}^A_\Phi$.*

The ability for an agent to construct answers inconsistent with her knowledge relies on the use of an X-inference in the definition of an answer (definition 7), and precisely on the fact that an agent $[K, X]$ can use a formula from $X \cap \overline{K}$. Indeed, $K \nvdash_X A$ iff $\exists x \in X \setminus \overline{K},\ K \cup A \vdash x$. But a given answer $R = K' \cup \text{conceivable}(X')$ of this agent to A only satisfies $K' \nvdash_{\{\bot\}\cup X'} A$, that is: $\exists x \in (\{\bot\} \cup X') \setminus \overline{K'},\ K' \cup A \vdash x$. Hence a formula of $X \cap \overline{K} \setminus K'$ is the reason why $K \nvdash_X A$, which explains why the answer is inconsistent with the knowledge of the agent.

Property 19. *Let M be a relevant answer of the agent $[K, X]$ to the set A of formulas. M is a lie of this agent regarding A iff there exists a formula of $M \setminus K$ inconsistent with K.*

Proof. ($\Rightarrow$) Since $M \in \mathcal{RL}^A_{[K,X]}$, $K \cup (M \setminus K) \vdash \bot$. It has been proved in (Aubry & Risch 2005) that for any answer R of $[K, X]$, $R \setminus K$ is either empty or a singleton. Hence $\exists m \in M \setminus K,\ K \cup \{m\} \vdash \bot$. ($\Leftarrow$) By assumption, $\exists m \in M \setminus K,\ K \cup \{m\} \vdash \bot$. Hence $K \cup M \vdash \bot$. From definition 18, it follows that $M \in \mathcal{RL}^A_{[K,X]}$. □

Corollary 20. *For every agent $[K, X]$ and for every set of formulas A, if the intersection of X and the deductive closure of K is empty, then this agent cannot generate any lie regarding A: $\forall[K, X], \forall A, X \cap \overline{K} = \varnothing \Rightarrow \mathcal{RL}^A_{[K,X]} = \varnothing$*

Proof. We prove the contrapositive, that is that $\mathcal{RL}^A_{[K,X]} \neq \varnothing \Rightarrow X \cap \overline{K} \neq \varnothing$. Assume $M \in \mathcal{RL}^A_{[K,X]}$. From property 19, $\exists m \in M \setminus K,\ K \cup \{m\} \vdash \bot$. Then $\exists m \in \text{conceivable}(X), K \vdash \neg m$. Hence $\exists x \in X, K \vdash x$, *i.e.* $X \cap \overline{K} \neq \varnothing$. □

This corollary leads to the conclusion that an agent with no knowledge ($[\varnothing, X]$), or whose way of reasoning is reduced to sole consistency ($[K, \{\bot\}]$) cannot generate any lie. Note however that a relevant answer of the agent $[\varnothing, X]$ to a set A can be a lie regarding A for an agent with the same set of forbidden formulas but with an adequate knowledge base:

Example 21. Let both Φ and Ψ be two agents such that $\Phi = [\varnothing, \{\bot, a\}]$, and $\Psi = [\{a\}, \{\bot, a\}]$. With $A = \{a\}$ we get: $\mathcal{RL}^A_\Phi = \varnothing$, and $\mathcal{R}r^A_\Phi = \{\{\neg a\}\}$, but: $\mathcal{RL}^A_\Psi = \{\{\neg a\}\}$.

A lie can be generated for attacking an argument while defending another argument:

Example 22 (Lie, attack, and defense)**.** Let Φ be an agent such that: $\Phi = [\{a, a \Rightarrow (b \wedge c)\}, \{\bot, a \wedge b\}]$. An agent Ψ set an argument α such that $\{b\} \in \text{elements}(\alpha)$. Then Φ could attack α while generating a lie: $\beta = \langle\{a, \neg(a \wedge b)\}, \neg b\rangle \in \mathcal{A}rg^{\text{att}(\alpha)}_\Phi$, and $\{a, \neg(a \wedge b)\} \in \mathcal{RL}^{\{b\}}_\Phi$. But if Ψ had set an argument α' with $\{\neg b\} \in \text{elements}(\alpha')$, then Φ could have defended α' with both the same lie and the same argument β: $\beta \in \mathcal{A}rg^{\text{def}(\alpha')}_\Phi$.

Finally, suppose that an agent Φ generates an argument α constructed from a lie, for answering an argument β. If $\alpha \in \mathcal{A}rg^{\text{def}(\beta)}_\Phi$ (resp. $\alpha \in \mathcal{A}rg^{\text{att}(\beta)}_\Phi$), we cannot presume whether there exists a coherent argument (*i.e.* not a lie) of Φ attacking β (resp. defending β). In other words we cannot assume that Φ is against (resp. for) any element of β:

Example 23 (Lie and neutral attitude)**.** Let Φ be an agent such that: $\Phi = [\{a\}, \{\bot, \neg b \vee a\}]$. Now consider an argument α such that $A = \{\neg b\}$ belongs to elements (α). Then Φ is neutral about A, but she could set up a counterargument to α, as far as it is constructed from a lie: $\langle\{\neg(\neg b \vee a)\}, b\rangle$.

Coming back to example 17, conferring the commercial agent the ability to lie allows her to push further her initial argument.

Example 24 (continuation of example 17)**.** Consider a commercial agent Φ' who still wants to sell a scooter of trademark *label-Z* to a client Ψ. $\Phi' = [\{scooter, label\text{-}Z, scooter \Rightarrow edgeOut, edgeOut \Rightarrow \neg trafficJam\}, \{\bot, (\boldsymbol{scooter} \wedge \boldsymbol{\neg label\text{-}Z}) \Rightarrow \boldsymbol{\neg trafficJam}\}]$.

Φ' set up the following argument: $\langle\{scooter \Rightarrow edgeOut, edgeOut \Rightarrow \neg trafficJam\}, (scooter \wedge label\text{-}Z) \Rightarrow \neg trafficJam\rangle$. The client address the following argument to the commercial agent: $\langle\{scooter \Rightarrow edgeOut, edgeOut \Rightarrow \neg trafficJam\}, scooter \Rightarrow \neg trafficJam\rangle$. Just like Φ in the example 17, the agent Φ' is for each element of this argument, *but* she can now generate a counterargument: $\langle\{\neg((\boldsymbol{scooter} \wedge \boldsymbol{\neg label\text{-}Z}) \Rightarrow \boldsymbol{\neg trafficJam})\}, \neg(scooter \Rightarrow \neg trafficJam)\rangle$.

Lies can be advantageous in negotiation dialogues when a goal is to be achieved. But it is possible to counterbalance this. Since the work of the philosopher Hamblin (Hamblin 1970), formal dialogue systems typically establish and maintain public sets of commitments called *commitment stores* for each agent. More than one notion of commitment is present in the literature on dialogues games [3] but essentially they can bee seen as a tool of memorization for the arguments advanced by each agent. If an agent Ψ considers an argument α from an agent Φ as a diluted argument (such as for instance in examples 17 and 24), then Ψ can question Φ about one or the other of the elements of α, and hope either a clarification (in example 17 where finally Φ agree with Ψ), or detect an inconsistency in the commitment store of Φ (in example 24 where Φ' has advanced two arguments whose supports are inconsistent together).

Conclusion

X-logics allows agents to cope with singular answers, namely those ones allowing the agent to produce an argument inconsistent with her knowledge, while keeping consistency in the knowledge base. This ability relies on the possibility for the agent to use expectations (if any) that contradict her knowledge. This way, a formal notion of lie is defined that seems to match the standard intuition. Moreover, a very simplified form of commitment store allows to define a formal way for detecting such lies. Of course, many questions still deserve to be addressed. A first question may concern how our system could be embedded in standard models of dialogue. Especially, having defined how agents may cheat, we are now concerned with the notion of strategy of an argumentative exchange. In this respect, it might be interesting to see how our system could be used for the formalization of some of the strategies described in (Schopenhauer 2004). Another may be more exciting question in the same direction concerns how an agent can exchange arguments with herself, thus allowing the representation of a form of introspection. Finally, some technical questions are still unanswered or under study, such as a link between the four possible attitudes of an agent with Belnap's four valued logic, or the comparison with the three *attitudes* described in (Parsons, Wooldridge, & Amgoud 2003).

[3] For recent works on the question, see for instance (Maudet & Chaib-draa 2003; Bentahar *et al.* 2004).

Acknowledgments

The authors are grateful to the anonymous reviewer for their helpful comments.

References

Amgoud, L., and Cayrol, C. 1998. On the acceptability of arguments in preference-based argumentation. In *UAI*, 1–7.

Amgoud, L., and Parsons, S. 2002. Agent dialogues with conflicting preferences. In *ATAL '01: Revised Papers from the 8th International Workshop on Intelligent Agents VIII*, 190–205. London, UK: Springer-Verlag.

Aubry, G., and Risch, V. 2005. Toward a logical tool for generating new arguments in an argumentation based framework. In *ICTAI'05*. Hong Kong, China: IEEE Computer Society. ISBN: 0–7695-2488–5.

Bench-Capon, T. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13(3):429–448.

Bentahar, J.; Moulin, B.; Meyer, J.-J. C.; and Chaib-draa, B. 2004. A logical model for commitment and argument network for agent communication. In *AAMAS '04: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, 792–799. Washington, DC, USA: IEEE Computer Society.

Besnard, P., and Hunter, A. 2001. A logic-based theory of deductive arguments. *Artificial Intelligence* 128(1-2):203–235.

Bochman, A. 2003. Brave nonmonotonic inference and its kinds. *Annals of Mathematics and Artificial Intelligence* 39(1–2):101–121.

Bondarenko, A.; Dung, P.; Kowalski, R.; and Toni, F. 1997. An abstract, argumentation-theoretic framework for default reasoning. *Artificial Intelligence* 93(1-2):63–101.

Dung, P. M.; Kowalski, R. A.; and Toni, F. 2006. Dialectic proof procedures for assumption-based, admissible argumentation. *Artificial Intelligence* 170(2):114–159.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77(2):321–358.

Elvang-Gøransson, M.; Krause, P.; and Fox, J. 1993. Dialectic reasoning with inconsistent information. In Heckerman, D., and Mamdani, E. H., eds., *UAI '93, The Catholic University of America, Providence, Washington, DC, USA*, 114–121. Morgan Kaufmann.

Hamblin, C. L. 1970. *Fallacies*. Methuen.

Lin, F., and Shoham, Y. 1989. Argument systems: A uniform basis for nonmonotonic reasoning. In Kaufmann, M., ed., *KR'89*, 245–255.

Maudet, N., and Chaib-draa, B. 2003. Commitment-based and dialogue-game based protocols – new trends in agent communication language. *Knowledge Engineering Review* 17(2):157–179.

Parsons, S.; Wooldridge, M.; and Amgoud, L. 2003. Properties and complexity of some formal inter-agent dialogues. *J. Log. Comput.* 13(3):347–376.

Prakken, H., and Vreeswijk, G. 2002. Logical systems for defeasible argumentation. In Gabbay, D., and Guenthner, F., eds., *Handbook of Philosophical Logic*, volume 4. Dordrecht: Kluwer Academic, 2nd edition. 219–318.

Prakken, H. 2005. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation* 15:1009–1040.

Rescher, N. 1977. *Dialectics, A Controversy-Oriented Approach to the Theory of Knowledge*. State University of New York Press, Albany.

Schopenhauer, A. 2004. *The Art Of Controversy*. Kessinger Publishing. translated by T. Bailey Saunders.

Siegel, P., and Forget, L. 1996. A representation theorem for preferential logics. In *KR'96*, 453–460. Morgan Kaufmann.

Simari, G. R., and Loui, R. P. 1992. A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence* 53(2-3):125–157.

Vreeswijk, G. 1992. Reasoning with defeasible arguments: Examples and applications. In Pearce, D., and Wagner, G., eds., *Logics in AI: Proc. of the European Workshop JELIA'92*. Berlin, Heidelberg: Springer. 189–211.