Vol. 00, No. 00, Month 2013, 1–26

RESEARCH ARTICLE

# An argumentation-based approach for reasoning about trust in information sources

Leila Amgoud Robert Demolombe

*IRIT – CNRS*
*118, route de Narbonne, 31062 Toulouse Cedex 9, FRANCE*

(*Received 00 Month 200x; final version received 00 Month 200x*)

During a dialog, agents exchange information with each other and need thus to deal with incoming information. For that purpose, they should be able to reason effectively about *trustworthiness* of information sources. This paper proposes an argument-based system that allows an agent to reason about its own beliefs and information received from other sources. An agent's beliefs are of two kinds: beliefs about the environment (like the window is closed) and beliefs about trusting sources (like agent $i$ trusts agent $j$). Six basic forms of trust are discussed in the paper including the most common one on sincerity. Starting with a base which contains such information, the system builds two types of arguments: arguments in favor of trusting a given source of information and arguments in favor of believing statements which may be received from other agents. We discuss how the different arguments interact and how an agent may decide to trust another source and thus to accept information coming from that source. The system is then extended in order to deal with graded trust (like agent $i$ trusts to some extent agent $j$).

**Keywords:** Trust, Argumentation, Modal Logic.

## 1. Introduction

An increasing number of software applications are being conceived, designed, and implemented using the notion of autonomous agents. These applications vary from email filtering Maes (1996), through electronic commerce Rodriguez et al. (1997), Wellman (1993), to large industrial applications Jennings et al. (1996). In all of these disparate cases, the agents are autonomous in the sense that they have the ability to decide for themselves which goals they should adopt and how these goals should be achieved Wooldridge and Jennings (1995). In most such applications, the autonomous components need to interact with one another because of the inherent interdependencies which exist between them. They need to communicate in order to resolve differences of opinion and conflicts of interest that result from differences in preferences, work together to find solutions to dilemmas and to construct proofs that they cannot manage alone, or simply to inform each other of pertinent facts. In other words they need the ability to engage in *dialogs*. Consequently, agents should be able to manage and deal with *trust* in information sources. In negotiation dialogs, for instance, one makes contracts with trustworthy agents. More generally, agents consider information coming from other sources only if these latter are trustworthy. As a result of this requirement on providing agents with the ability to

ISSN: 1946-2166 print/ISSN 1946-2174 online

DOI: 10.1080/1946216YYxxxxxxxx
http://www.informaworld.com

deal with trust, an important amount of work has been done. Two main categories of works can be distinguished:

- Works on understanding and formalizing the notion of trust in information sources. Such works try to answer the question: what does the sentence "agent $x$ trusts agent $y$" mean? Examples of answers can be found in Castelfranchi (2011), Castelfranchi and Falcone (2000), Falcone et al. (2013), Marsh (1994). In Demolombe (1998, 1999), it is argued that trust is generally not absolute but rather concerns some properties of an agent like his *competence*, *sincerity*, *cooperativity*, ....
- Works on reasoning about trust. The idea is to decide whether to trust or not a given source of information. Two categories of models are particularly proposed: i) statistics-based models (e.g., Matt et al. (2010), Shi et al. (2005)) which rely on past behavior of a source in order to predict its future behavior. ii) logical models (e.g., Demolombe (2004), Demolombe and Lorini (2008)) which infer trust in some properties from trust in other properties.

Besides, since the seminal book by Walton and Krabbe (1995) in which they distinguished between six types of dialogs, there has been much work on providing agents with the ability to engage in such dialogs. Typically, these focus on one type of dialog like persuasion (e.g. Amgoud et al. (2000)), inquiry (e.g. Black and Hunter (2009)), negotiation (e.g. Sycara (1990)) and deliberation (e.g. McBurney et al. (2007)). Furthermore, Walton and Krabbe emphasized the need to argue in dialogs in order to convince other parties to accept opinions or offers. Consequently, in most works on modeling dialogs, agents are equipped with argumentation systems for reasoning about their own beliefs, building arguments and evaluating arguments received from other sources. While this use of argumentation is a common theme in all work mentioned above, none of those proposals consider trust in information sources when dealing with incoming information or when making deals with other agents. They rather assume that agents are trustworthy and accept any information (respectively offer) sent by any agent as soon as it does not contradict their own beliefs (respectively, it satisfies their goals). However, agents are not necessarily neither sincere nor reliable as argued in the huge literature about trust in information sources. This would mean that in existing works, agents may accept claims even if their sources are not trustworthy. They may also make deals with unreliable agents.

This paper fills the gap by proposing an argumentation system that agents may use in dialogs for reasoning about different kinds of beliefs including beliefs about trust in information sources. The system fulfills thus three tasks. It states whether:

- to believe in a given statement,
- to trust or not a given source,
- to accept or not an information/offer received from a source.

We consider a fine-grained notion of trust as opposed to absolute trust. Indeed, an agent trusts (or distrusts) another agent in a given property and not in absolute way. For instance, one may trust someone is his sincerity but not in his competence. In this paper, we focus on the six properties identified by Demolombe (1998, 2004), namely *validity*, *completeness*, *sincerity*, *cooperativity*, *competence* and *vigilance*. In the first part of the paper, trust is considered as a binary notion, i.e., an agent either trusts in a given property of an entity or not. The system

starts with a belief base which is encoded in modal logic and which contains formulas expressing information about the environment (e.g., my car is red) and information about trust (e.g., agent $i$ trusts in the sincerity of agent $j$). It builds arguments in favor of statements and establishes the attacks between them. The arguments are evaluated using Dung's semantics (Dung (1995)), and finally the inferences to be drawn from the base are identified. We show that the system satisfies nice properties, namely the rationality postulates defined in Amgoud (2013) about consistency and closure under consequence operator. In the second part of the paper, the system is extended in order to deal with graded trust as developed in Demolombe (2009) and in Demolombe and Liau (2001). The logical language that is used for representing beliefs is extended in such a way to encode *certainty* degrees of beliefs (like, agent i has some doubts about climate change) and *regularities* degrees of relationships between facts (like, if we are in London, it is raining almost every day). From these two kinds of degrees, each argument is assigned an importance level which may not be the same for all arguments. Finally, arguments are evaluated using not only the attack relation but also a preference relation issued from the importance levels of arguments.

The paper is structured as follows: Section 2 introduces the logical formalism that will be used for representing and reasoning about agent's beliefs. Section 3 defines the six forms of trust that were initially introduced in Demolombe (2004), Lorini and Demolombe (2008) in case of binary trust. Section 4 presents the argumentation system as well as it properties. Section 5 presents the graded version of trust as proposed in Demolombe (2009), Demolombe and Liau (2001), and an argumentation system that can take into account varying degrees of trust and beliefs. Section 6 compares our model with existing works on argumentation-based trust. The last section concludes.

## 2. Logical formalism

This section introduces the logical framework (i.e., the logical language $\mathcal{L}$ and its axiomatics) that will be used for representing and reasoning about beliefs and trust in information sources. The syntactic primitives of $\mathcal{L}$ are the following:

- ATOM: set of atomic propositions denoted by $p$, $q$, $r, \ldots$
- AGENT: a non-empty set of agents denoted by $i$, $j$, $k$, $\ldots$

The language $\mathcal{L}$ is the set of formulas defined by the following BNF:

$$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid \mathrm{Bel}_i\phi \mid \mathrm{Inf}_{j,i}\phi$$

where $p$ ranges over ATOM and $i$ and $j$ range over AGENT. The other logical connectives are defined as usual. The intuitive meaning of the modal operators is:

- $\mathrm{Bel}_i\phi$[1]: agent $i$ believes that $\phi$ holds
- $\mathrm{Inf}_{j,i}\phi$: agent $j$ has informed agent $i$ that $\phi$ holds

[1]Sometimes we abuse notation and write $\mathrm{Bel}_i(\phi)$ instead of $\mathrm{Bel}_i\phi$.

The axiomatics of the logic is the axiomatics of a Propositional multi Modal Logic (Chellas (1980)). Indeed, in addition to the axiomatics of Classical Propositional Calculus we have the following axiom schemas and inference rules.

(K) $\text{Bel}_i(\phi \rightarrow \psi) \rightarrow (\text{Bel}_i\phi \rightarrow \text{Bel}_i\psi)$
(D) $\neg(\text{Bel}_i\phi \wedge \text{Bel}_i\neg\phi)$
(Nec) If $\vdash \phi$, then $\vdash \text{Bel}_i\phi$

Roughly speaking the intuitive meaning of (K) is that agent $i$ can apply the *modus ponens* rule to derive consequences, (D) means that $i$'s beliefs are not inconsistent and (Nec) means that $i$ is not ignorant of the logical truths.

The modal operator $\text{Inf}_{j,i}$ obeys the following axiom schemas:

(EQV) If $\vdash \phi \leftrightarrow \psi$, then $\vdash \text{Inf}_{j,i}\phi \leftrightarrow \text{Inf}_{j,i}\psi$
(CONJ) $\text{Inf}_{j,i}\phi \wedge \text{Inf}_{j,i}\psi \rightarrow \text{Inf}_{j,i}(\phi \wedge \psi)$
(OBS) $\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_i\text{Inf}_{j,i}\phi$
(OBS') $\neg\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_i\neg\text{Inf}_{j,i}\phi$

The intuitive meaning of (EQV) is that informing actions about two logically equivalent formulas have the same effects. For instance, to inform about the fact John is at home and John is working has the same effects as to inform about the fact that John is working and John is at home. The meaning of (CONJ) is that to inform about the fact John is at home and to inform about the fact John is working has the same effects as to inform about the fact John is working at home. The justification of this axiom schema is that informing actions are considered at an abstract level and two distinct concrete actions may be considered as the "same" action if they produce the same effect on the receiver's beliefs. The axiom schemas (OBS) and (OBS') assume that if an agent $j$ informs (respectively does not inform) an agent $i$ about $\phi$, then $i$ is aware of this fact. This would mean that the communication channels are assumed to be perfect.

According to Chellas's terminology, modalities such as $\text{Bel}_i$ obey a normal system KD and modalities of the kind $\text{Inf}_{j,i}$ obey a particular kind of classical system. Axiom schemas (OBS) and (OBS') show how these two kinds of modalities interact.

In the sequel, the symbol $\vdash$ refers to the consequence operator that is based on the previous axiom schemas. Besides, a *belief base* is a subset of $\mathcal{L}$ which contains the beliefs of a given agent $i \in \text{AGENT}$.

## 3. Binary trust in information sources

Throughout this section, we consider two interacting agents $i$ and $j$ and assume that $i$ receives a piece of information $\phi \in \mathcal{L}$ from agent $j$. An important question is then what is the effect of this action on what the receiver believes? In Demolombe (1998, 2004), it was argued that this depends on the sender's *properties* the receiver trusts in. Six properties were particularly distinguished and investigated:

**Trust in *sincerity*:** *sincerity* is the relationship between what the trustee says and what he believes. For instance, the fact that Juliet trusts Romeo in his sincerity about the fact Juliet is beautiful means that Juliet believes that if Romeo says to Juliet that she is beautiful, then Romeo believes that she is beautiful. The general

definition is: the truster believes that if he is informed by the trustee about some proposition, then the trustee believes that this proposition is true. Formally:

$$\text{TrustSinc}(i, j, \phi) \stackrel{\text{def}}{=} \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi)$$

It is worth mentioning that the fact that an agent $i$ believes in the sincerity of another agent $j$ regarding proposition $\phi$ does not mean that $i$ believes $\phi$. The claim may be false and $j$ is not aware about that. A strong version of sincerity is the property of validity.

**Trust in *validity*:** *validity* is the relationship between what the trustee says and what is true. For instance, the fact that Romeo trusts Juliet in her validity about the fact that Juliet loves Romeo means that Romeo believes that if Juliet says to Romeo that she loves him, then it is true that she loves him. The general definition is: the truster ($i$) believes that if he is informed by the trustee ($j$) about some proposition, then this proposition is true.

$$\text{TrustVal}(i, j, \phi) \stackrel{\text{def}}{=} \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \phi)$$

**Trust in *completeness*:** *completeness* is the relationship between what is true and what the trustee says; it is the dual of *validity*. For instance the fact that Romeo trusts Juliet in her completeness about the fact that Juliet loves Romeo means that Romeo believes that if it is true that Juliet loves him, then Juliet will tell Romeo that she loves him. The general definition is: the truster believes that if some proposition is true, then he is informed by the trustee about this proposition.

$$\text{TrustCmp}(i, j, \phi) \stackrel{\text{def}}{=} \text{Bel}_i(\phi \rightarrow \text{Inf}_{j,i}\phi)$$

**Trust in *cooperativity*:** *cooperativity* is the relationship between what the trustee believes and what he says; it is the dual of sincerity. For instance, the fact that Juliet trusts Romeo in his cooperativity about the fact Juliet is beautiful means that Juliet believes that if Romeo believes that she is beautiful, then Romeo says to her that she is beautiful. The general definition is: the truster believes that if the trustee believes that some proposition is true, then he is informed by the trustee about this proposition.

$$\text{TrustCoop}(i, j, \phi) \stackrel{\text{def}}{=} \text{Bel}_i(\text{Bel}_j\phi \rightarrow \text{Inf}_{j,i}\phi)$$

**Trust in *competence*:** *competence* is the relationship between what the trustee believes and what is true. For instance, the fact that Juliet trusts Romeo in his competence about the fact that the door of her house is closed means that Juliet believes that if Romeo believes that the door of her house is closed, then it is true that the door is closed. The general definition is: the truster believes that if the trustee believes that some proposition is true, then this proposition is true.

$$\text{TrustComp}(i, j, \phi) \stackrel{\text{def}}{=} \text{Bel}_i(\text{Bel}_j\phi \rightarrow \phi)$$

**Trust in *vigilance*:** *vigilance* is the relationship between what is true and what the trustee believes; it is the dual of competence. For instance, the fact that Juliet trusts Romeo in his vigilance about the fact that the door of her house is closed

means that Juliet believes that if it is true that the door of her house is closed, then Romeo believes that the door of her house is closed. The general definition is: the truster believes that if some proposition is true, then the trustee believes that this proposition is true.

$$\text{TrustVigi}(i, j, \phi) \stackrel{\text{def}}{=} \text{Bel}_i(\phi \rightarrow \text{Bel}_j\phi)$$

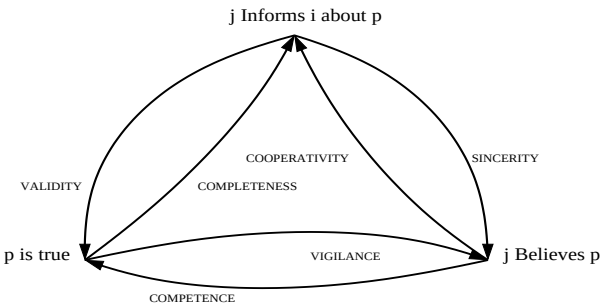



Figure 1. Relationships between believing, informing and truth.

In Parsons et al. (2012) other properties, called *argument schemes*, are discussed like trust in agent's reputation or trust in agent's character. For the purpose of the paper, we only focus on the six above properties and propose a formal framework for reasoning with and about them.

**Remarks:** It is worth mentioning that the presented definitions of trust are specific to particular propositions. For instance, a patient ($p$) may trust in the competence of his doctor ($d$) regarding diagnosis $g_1$. This is represented by the formula $\text{Bel}_p(\text{Bel}_d g_1 \rightarrow g_1)$. This does not mean that the patient trusts also his doctor on another diagnosis $g_2$. Note also that the six formulas are elements of $\mathcal{L}$.

As said before, completeness is the dual of validity, cooperativity is the dual of sincerity and vigilance is the dual of competence (see Figure 1). The dual properties play a significant role. Let us consider the case where the trustee is a guard in charge of informing people living in a building if the elevator fails. If these people trust the guard in his completeness, they infer that the elevator is working from the fact they have not received a warning from the guard.

It is also easy to show that the six properties are not independent. Indeed, trust in validity follows from trust in sincerity and trust in competence. Similarly, trust in completeness follows from trust in vigilance and trust in cooperativity. In formal terms we have:

(V) $\quad \vdash \text{TrustSinc}(i, j, \phi) \wedge \text{TrustComp}(i, j, \phi) \rightarrow \text{TrustVal}(i, j, \phi)$
(C) $\quad \vdash \text{TrustVigi}(i, j, \phi) \wedge \text{TrustCoop}(i, j, \phi) \rightarrow \text{TrustCmp}(i, j, \phi)$

The effects of informing actions depending on the different kinds of trust are summarized below:

(E1) $\quad \vdash \text{TrustSinc}(i, j, \phi) \rightarrow (\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_i\text{Bel}_j\phi)$
(E2) $\quad \vdash \text{TrustVal}(i, j, \phi) \rightarrow (\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_i\phi)$

(E3) $\vdash \text{TrustCoop}(i,j,\phi) \rightarrow (\neg \text{Inf}_{j,i}\phi \rightarrow \text{Bel}_i \neg \text{Bel}_j \phi)$

(E4) $\vdash \text{TrustCmp}(i,j,\phi) \rightarrow (\neg \text{Inf}_{j,i}\phi \rightarrow \text{Bel}_i \neg \phi)$

Property (E2) (resp. (E4)) shows sufficient conditions about trust that guarantee that performing (resp. not performing) the action $\text{Inf}_{j,i}\phi$ has the effect that $i$ believes that $\phi$ is true (resp. false). Notice that from $i$'s trust in $j$ competence (resp. trust vigilance) performing (resp. not performing) the action $\text{Inf}_{j,i}\phi$ does not allow $i$ to infer that $\phi$ is true (resp. false). For instance, even if $i$ trusts the doctor $j$ about his competence about cancer diagnosis, $i$ may not trust him about his sincerity, and if the doctor tells him that he has no cancer, $i$ will not believe that he has not a cancer. The reason why $i$ does not trust the doctor about his sincerity may be that $i$ believes that the doctor wants to protect $i$ from bad news.

The effects of informing actions can be derived from the different kinds of assumptions about the trust relationships between agents. For instance, if the truster $i$ trusts $j$ in his sincerity about the proposition $\phi$ and $j$ informs $i$ about $\phi$, the truster can infer that the trustee believes what he has transmitted to him ($i$). If, in addition, the truster trusts $j$ in his competence (i.e., the formula $\text{Bel}_i(\text{Bel}_j\phi \rightarrow \phi)$ is in the beliefs base of agent $i$), then the truster can infer that $\phi$ is true. Notice that this consequence is in the scope of what the truster believes (i.e., what is inferred is $\text{Bel}_i\phi$ and not $\phi$). Let us assume, for instance, that the truster $i$ has some disease, $j$ is a doctor and $j$ tells to $i$ that $i$ has a flu. If $i$ trusts his doctor in his sincerity about this diagnosis, $i$ can infer that the doctor does believe that $i$ has a flu. If $i$ also trusts his doctor about his competence, $i$ can infer that he has a flu. Then, the final effects of what the doctor said is that $i$ believes that the doctor believes that $i$ has a flu and also that $i$ believes that he has a flu. Notice that, if $i$ trusts the doctor only in his validity, the effect of what the doctor said is that $i$ believes that he has a flu but it is not necessarily the case that $i$ believes that the doctor believes that $i$ has a flu (seeDemolombe (2011)). Indeed, it could be the case that $i$ believes that the doctor just transmits a diagnosis that has been made by an assistant who is trusted to be sincere and competent, while the doctor is not. From a formal point of view, it is not necessarily the case that contraposition of property (V) holds.

## 4. Argumentation-based reasoning system

Argumentation is seen as a reasoning process in which arguments are built and evaluated in order to increase or decrease the acceptability of a given standpoint. The latter may be a belief, an action, a goal, etc. Argumentation has become an Artificial Intelligence keyword for the last twenty years. In its essence, argumentation can be seen as a particularly useful and intuitive paradigm for doing nonmonotonic reasoning. The advantage of argumentation is that the reasoning process is composed of modular and quite intuitive steps, and thus avoids the monolithic approach of many traditional logics for defeasible reasoning. An argumentation process starts with the construction of a set of arguments from a given knowledge base. As some of these arguments may attack each other, one needs to apply a criterion for determining the sets of arguments that can be regarded as acceptable: the so-called *extensions*.

In what follows, we propose an argumentation system for reasoning about the different kinds of beliefs an agent $i$ may have, in particular beliefs about trust in information sources. The system instantiates the abstract framework of Dung (1995) and uses one of its semantics in order to evaluate arguments. Before presenting

the system, we start by recalling briefly Dung's framework and then show how arguments in favor of beliefs can be built and how these arguments may interact with each other.

### 4.1. *Dung's abstract argumentation framework*

The most abstract *argumentation framework* in the literature was proposed by Dung (1995). It consists of a set of *arguments* and a binary relation expressing *attacks* between the arguments. Both notions (i.e., arguments and attacks) are abstract entities and thus, their origin and structure are left unspecified.

Definition 4.1 An *argumentation framework* is a pair $(\mathcal{A}, \mathcal{R})$ where $\mathcal{A}$ is a set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is an attack relation.

A pair $(a, b) \in \mathcal{R}$ means that $a$ attacks $b$. A set $\mathcal{E} \subseteq \mathcal{A}$ attacks an argument $b$ iff $\exists a \in \mathcal{E}$ such that $(a, b) \in \mathcal{R}$. We sometimes use the infix notation $a\mathcal{R}b$ to denote $(a, b) \in \mathcal{R}$.

An argumentation framework $(\mathcal{A}, \mathcal{R})$ is seen as a *graph* whose nodes are the arguments of $\mathcal{A}$ and its edges are the attacks in $\mathcal{R}$. The arguments are evaluated using a *semantics*. In Dung (1995), different semantics were proposed, and some of them were refined, for instance in Baroni et al. (2005), Dung et al. (2007). For the purpose of the paper, we only recall *stable* semantics since our aim is not to discuss the outcomes of our system under all semantics, but rather to show how to build arguments in favor of trust in information sources and how to decide to accept information coming from sources. Thus, we only need one semantics for illustration purposes.

Definition 4.2 Let $\mathcal{T} = (\mathcal{A}, \mathcal{R})$ be an argumentation framework and $\mathcal{E} \subseteq \mathcal{A}$. $\mathcal{E}$ is a *stable* extension iff:

- $\nexists a, b \in \mathcal{E}$ such that $(a, b) \in \mathcal{R}$
- $\mathcal{E}$ attacks any argument in $\mathcal{A} \setminus \mathcal{E}$

$\text{Ext}(\mathcal{T})$ denotes the set of all stable extensions of $\mathcal{T}$.

It is worth recalling that stable extensions are maximal (for set inclusion) non-conflicting sets of arguments.

Example 4.3 Let us consider the argumentation framework $\mathcal{T} = (\mathcal{A}, \mathcal{R})$ such that:

- $\mathcal{A} = \{a, b, c, d, e, f, g\}$
- $\mathcal{R} = \{(c, b), (b, e), (e, c), (d, c), (a, d), (d, a), (a, f), (f, g)\}$

This framework has five maximal (for set inclusion) non-conflicting sets of arguments:

- $\mathcal{E}_1 = \{a, c, g\}$,
- $\mathcal{E}_2 = \{d, e, f\}$,
- $\mathcal{E}_3 = \{b, d, f\}$,
- $\mathcal{E}_4 = \{a, e, g\}$, and
- $\mathcal{E}_5 = \{a, b, g\}$.

It has one stable extension $\mathcal{E}_3$, i.e., $\text{Ext}(\mathcal{T}) = \{\mathcal{E}_3\}$.

An argumentation framework may be *infinite*, i.e., its set of arguments may be infinite. Consequently, it may have an infinite number of extensions (under a given semantics).

### 4.2. *Binary trust supported by arguments*

This section introduces an argumentation system for reasoning about the different kinds of beliefs an agent $i$ may have. As already said, argumentation is an alternative approach for reasoning with inconsistent information. It follows three main steps: i) constructing *arguments* and counterarguments from a logical belief base, ii) defining the *status* of each argument, and iii) specifying the *conclusions* to be drawn from the base. In what follows, we focus on a given agent $i$ and propose a model for reasoning about his beliefs. The model instantiates Dung's framework by defining all the above items.

Starting from the logic $(\mathcal{L}, \vdash)$ described in Section 2 and a possibly inconsistent beliefs base $\mathcal{K}_i \subseteq \mathcal{L}$, the system computes a consistent set of beliefs the agent should rely on. The base $\mathcal{K}_i$ can be seen as the $i$'s "candidate" beliefs. It may contain trust information as defined in the previous section (e.g., $\text{Bel}_i(\phi \rightarrow \text{Bel}_j\phi)$), beliefs about the environment (e.g., $\text{Bel}_i\phi$ where $\phi$ stands for 'the window is closed') and beliefs about informing actions received from other agents (e.g., $\text{Bel}_i\text{Inf}_{j,i}\phi$). Note that the base $\mathcal{K}_i = \{\text{Bel}_i\text{Inf}_{j,i}\phi, \text{Bel}_i\text{Inf}_{j,i}\neg\phi\}$ is not inconsistent. Here agent $i$ believes that he was informed by $j$ that both formulas $\phi$ and $\neg\phi$ hold. However, the base $\mathcal{K}_i = \{\text{Bel}_i\phi, \text{Bel}_i\neg\phi\}$ is inconsistent.

The system is a logical instantiation of the abstract framework proposed by Dung in his seminal paper Dung (1995). It consists thus of a set of arguments, an attack relation between the arguments and a semantics for evaluating the arguments. The arguments are built from the base $\mathcal{K}_i$. They are logical proofs for formulas in $\mathcal{L}$ that satisfy two requirements: consistency and minimality.

Definition 4.4 An *argument* built from a belief base $\mathcal{K}_i$ is a pair $(H, h)$ where:

- $H \subseteq \mathcal{K}_i$ and $h \in \mathcal{L}$
- $H$ is consistent
- $H \vdash h$
- $\nexists H' \subset H$ such that $H' \vdash h$

$H$ is called the *support* of the argument and $h$ its *conclusion*. $\text{Arg}(\mathcal{K}_i)$ is the set of all arguments that can be built from $\mathcal{K}_i$.

Let us illustrate this notion of argument with an example.

Example 4.5 Assume the following belief base of agent $i$:

$$\mathcal{K}_i = \begin{cases} \text{Bel}_i(\delta) \\ \text{Bel}_i(\text{Inf}_{j,i}\phi) \\ \text{Bel}_i(\neg\text{Inf}_{k,i}\varphi) \\ \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi) \\ \text{Bel}_i(\varphi \rightarrow \text{Inf}_{k,i}\varphi) \end{cases}$$

From $\mathcal{K}_i$, an infinite number of arguments are built including the following ones:

(1) $(\{\text{Bel}_i(\delta)\}, \text{Bel}_i(\delta))$
(2) $(\{\text{Bel}_i(\text{Inf}_{j,i}\phi)\}, \text{Bel}_i(\text{Inf}_{j,i}\phi))$
(3) $(\{\text{Bel}_i(\neg\text{Inf}_{k,i}\varphi)\}, \text{Bel}_i(\neg\text{Inf}_{k,i}\varphi))$
(4) $(\{\text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi), \text{Bel}_i(\text{Inf}_{j,i}\phi)\}, \text{Bel}_i(\text{Bel}_j\phi))$
(5) $(\{\text{Bel}_i(\varphi \rightarrow \text{Inf}_{k,i}\varphi), \text{Bel}_i(\neg\text{Inf}_{k,i}\varphi)\}, \text{Bel}_i\neg\varphi)$

The previous arguments support various beliefs of agent $i$. Some of them, like (4) and (5), make use of beliefs on trust in information sources. To put it differently, they rely on agent's trust in order to make inferences. Such arguments are very useful in dialog systems where agents may receive new information from other entities and should thus decide whether to accept it or not.

Arguments may also support the six forms of trust we discussed in Section 3. They show whether agent $i$ should or not trust another agent in one of the properties (sincerity, validity, cooperativity, completeness and competence). Let us consider the following example.

Example 4.6 Assume the following base:

$$\mathcal{K}_i = \begin{cases} \text{Bel}_i(\varphi) \rightarrow \text{TrustSinc}(i, j, \phi) \\ \text{TrustVal}(i, k, \varphi) \\ \text{Bel}_i(\text{Inf}_{k,i}\varphi) \end{cases}$$

where $i$ is the program chair of a conference, $k$ is an area chair member of the program committee and $j$ is a reviewer. Assume that $\varphi$ stands for "$j$ makes fair reviews" and $\phi$ for "$j$ makes a fair review for paper ID $x$". Examples of arguments that are built from this base are the following ones:

(1) $(\{\text{Bel}_i(\text{Inf}_{k,i}\varphi)\}, \text{Bel}_i(\text{Inf}_{k,i}\varphi))$
(2) $(\{\text{Bel}_i(\text{Inf}_{k,i}\varphi), \text{TrustVal}(i, k, \varphi)\}, \text{Bel}_i\varphi)$
(3) $(\{\text{Bel}_i(\text{Inf}_{k,i}\varphi),\ \text{TrustVal}(i, k, \varphi),$
$\text{Bel}_i\varphi \rightarrow \text{TrustSinc}(i, j, \phi)\},\ \text{TrustSinc}(i, j, \phi))$

Note that the argument (3) is in favor of trusting in the sincerity of agent $j$ regarding proposition $\phi$.

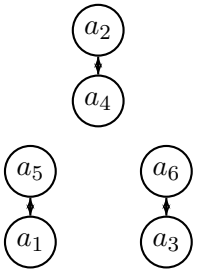
The second component of an argumentation framework is its attack relation which expresses conflicts that may raise between arguments. In argumentation literature, several relations were proposed (see Gorogiannis and Hunter (2011) for a summary of relations proposed for propositional frameworks). Some of them, like the well-known *rebutting*, are symmetric. However, it was shown in Amgoud and Besnard (2009) that any argumentation framework which is grounded on a Tarskian logic (Tarski (1956)) and uses a symmetric attack relation may violate the rationality postulates proposed in Caminada and Amgoud (2007), namely the one on consistency. Indeed, such a framework may have an extension which supports inconsistent conclusions. Since modal logic is a particular case of Tarski's logics, then the argumentation system we propose here will suffer from the same problem as shown in the following example.

Example 4.7 Let us consider the following belief base:

$$\mathcal{K}_i = \begin{cases} \text{Bel}_i(\phi) \\ \text{Bel}_i(\neg\varphi) \\ \text{Bel}_i(\phi \rightarrow \varphi) \end{cases}$$

Let us consider the following arguments:

- $a_1 = (\{\text{Bel}_i(\phi)\}, \text{Bel}_i(\phi))$
- $a_2 = (\{\text{Bel}_i(\neg\varphi)\}, \text{Bel}_i(\neg\varphi))$
- $a_3 = (\{\text{Bel}_i(\phi \rightarrow \varphi)\}, \text{Bel}_i(\phi \rightarrow \varphi))$
- $a_4 = (\{\text{Bel}_i(\phi), \text{Bel}_i(\phi \rightarrow \varphi)\}, \text{Bel}_i(\varphi))$
- $a_5 = (\{\text{Bel}_i(\neg\varphi), \text{Bel}_i(\phi \rightarrow \varphi)\}, \text{Bel}_i(\neg\phi))$
- $a_6 = (\{\text{Bel}_i(\phi), \text{Bel}_i(\neg\varphi)\}, \text{Bel}_i(\phi \wedge \neg\varphi))$




Let $\mathcal{R}$ be the *rebutting* relation defined as follows: $(H, h)$ *rebuts* $(H', h')$ iff $h = \text{Bel}_i\phi$, $h' = \text{Bel}_i\varphi$ and $\phi \equiv \neg\varphi$. Note that this relation is symmetric. The attacks among arguments are as depicted in figure above. The set $\{a_1, a_2, a_3\}$ is a stable extension of $(\text{Arg}(\mathcal{K}_i), \mathcal{R})$. However, $\{\text{Bel}_i(\phi), \text{Bel}_i(\neg\varphi), \text{Bel}_i(\phi \rightarrow \varphi)\}$ is inconsistent. This means that the extension supports contradictory conclusions!

In what follows we avoid thus symmetric relations. We discuss next various forms of attacks. The first one is the so-called *assumption-attack* proposed in Elvang-Gøransson et al. (1993). It consists of weakening an argument by undermining one of its premises (i.e., an element of its support).

Definition 4.8 Let $(H, h), (H', h')$ be two arguments of $\text{Arg}(\mathcal{K}_i)$. $(H, h)$ *assumption-attacks* $(H', h')$ iff there exists $h'' \in H'$ such that $h = \text{Bel}_i\phi$ and $h'' = \text{Bel}_i\neg\phi$.

Let us illustrate this relation on the following example.

Example 4.9 Let us consider the following base:

$$\mathcal{K}_i = \begin{cases} \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi) \\ \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \phi) \\ \text{Bel}_i(\text{Inf}_{j,i}\phi) \\ \text{Bel}_i(\neg\phi) \end{cases}$$

The argument $(\{\text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \phi), \text{Bel}_i(\neg\phi)\}, \text{Bel}_i(\neg\text{Inf}_{j,i}\phi))$ assumption attacks the argument $(\{\text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi), \text{Bel}_i(\text{Inf}_{j,i}\phi)\}, \text{Bel}_i(\text{Bel}_j\phi))$.

It is worth mentioning that this attack relation concerns all types of arguments that may be built from a beliefs base (i.e., arguments supporting ordinary beliefs and those supporting trust in information sources). The following definition introduces another way for attacking arguments in favor of trust in an agent's sincerity. The basic idea is to show a *case* where the trusted agent sent an information that he does not believe. To put it differently, the attack consists of proving that the trustee may lie.

Definition 4.10 Let $(H, h), (H', h')$ be two arguments of $\text{Arg}(\mathcal{K}_i)$. $(H, h)$ *sinc-attacks* $(H', h')$ iff $h = \text{Bel}_i(\text{Inf}_{j,i}\varphi \wedge \neg\text{Bel}_j\varphi)$ and $\text{TrustSinc}(i, j, \phi) \in H'$.

An argument in favor of trust in validity may also be undermined by an argument whose conclusion is a formula which is sent by the trusted agent and which is invalid (i.e., it does not hold).

Definition 4.11 Let $(H, h), (H', h')$ be two arguments of $\text{Arg}(\mathcal{K}_i)$. $(H, h)$ *val-attacks* $(H', h')$ iff $h = \text{Bel}_i(\text{Inf}_{j,i}\varphi \wedge \neg\varphi)$ and $\text{TrustVal}(i, j, \phi) \in H'$.

Similarly, an argument in favor of trust in completeness may be attacked. Recall that such an argument provides a reason for believing that if a given formula

holds, then the truster agent will be informed about it by the trustee. An attacker highlights a formula which holds and for which the trustee does not send any message.

Definition 4.12 Let $(H,h),(H',h')$ be two arguments of $\mathrm{Arg}(\mathcal{K}_i)$. $(H,h)$ *com-attacks* $(H',h')$ iff $h = \mathrm{Bel}_i(\varphi \wedge \neg \mathrm{Inf}_{j,i}\varphi)$ and $\mathrm{TrustCmp}(i,j,\phi) \in H'$.

Recall that trust in the cooperativity of an agent means that if he believes a statement, then he will inform the truster about it. An attack against an argument supporting such information consists of presenting a case where the trustee was not cooperative.

Definition 4.13 Let $(H,h),(H',h')$ be two arguments of $\mathrm{Arg}(\mathcal{K}_i)$. $(H,h)$ *coop-attacks* $(H',h')$ iff $h = \mathrm{Bel}_i(\mathrm{Bel}_j\varphi \wedge \neg \mathrm{Inf}_{j,i}\varphi)$ and $\mathrm{TrustCoop}(i,j,\phi) \in H'$.

An argument in favor of trust in the competence of an agent may be attacked by an argument supporting a statement that is believed by this agent but which is not true.

Definition 4.14 Let $(H,h),(H',h')$ be two arguments of $\mathrm{Arg}(\mathcal{K}_i)$. $(H,h)$ *comp-attacks* $(H',h')$ iff $h = \mathrm{Bel}_i(\mathrm{Bel}_j\varphi \wedge \neg\varphi)$ and $\mathrm{TrustComp}(i,j,\phi) \in H'$.

Trust in an agent's vigilance may be attacked by exhibiting a claim which holds but is ignored by the agent.

Definition 4.15 Let $(H,h),(H',h')$ be two arguments of $\mathrm{Arg}(\mathcal{K}_i)$. $(H,h)$ *vigi-attacks* $(H',h')$ iff $h = \mathrm{Bel}_i(\varphi \wedge \neg \mathrm{Bel}_j\varphi)$ and $\mathrm{TrustVigi}(i,j,\phi) \in H'$.

**Remark:** It is worth mentioning that assumption-attack relation is *conflict-dependent*, i.e., if $(H,h)$ attacks $(H',h')$ then $H \cup H'$ is necessarily inconsistent. This is not the case for the six other relations as shown in the following example.

Example 4.16 Let us consider the following base:

$$\mathcal{K}_i = \begin{cases} \mathrm{Bel}_i(\mathrm{Inf}_{j,i}\phi \rightarrow \mathrm{Bel}_j\phi) \\ \mathrm{Bel}_i(\mathrm{Inf}_{j,i}\varphi) \\ \mathrm{Bel}_i(\neg \mathrm{Bel}_j\varphi) \end{cases}$$

Assume that $\phi$ stands for 'The weather is cloudy' and $\varphi$ stands for 'People pay few taxes'. Note that the base $\mathcal{K}_i$ is consistent. However, the argument $(\{\mathrm{Bel}_i(\mathrm{Inf}_{j,i}\varphi), \mathrm{Bel}_i(\neg \mathrm{Bel}_j\varphi)\}, \mathrm{Bel}_i(\mathrm{Inf}_{j,i}\varphi \wedge \neg \mathrm{Bel}_j\varphi))$ sinc-attacks the argument $(\{\mathrm{Bel}_i(\mathrm{Inf}_{j,i}\phi \rightarrow \mathrm{Bel}_j\phi)\}, \mathrm{Bel}_i(\mathrm{Inf}_{j,i}\phi \rightarrow \mathrm{Bel}_j\phi))$.

The seven forms of attacks are captured by a binary relation on the set of arguments which is denoted by $\Re$.

Definition 4.17 Let $(H,h)$ and $(H',h')$ be two arguments of $\mathrm{Arg}(\mathcal{K}_i)$. $(H,h)$ $\Re$ $(H',h')$ iff:

- $(H,h)$ *assumption-attacks* $(H',h')$, or
- $(H,h)$ *sinc-attacks* $(H',h')$, or
- $(H,h)$ *val-attacks* $(H',h')$, or
- $(H,h)$ *com-attacks* $(H',h')$, or
- $(H,h)$ *coop-attacks* $(H',h')$, or
- $(H,h)$ *comp-attacks* $(H',h')$, or

- $(H, h)$ *vigi-attacks* $(H', h')$.

The following example shows that the attack relation $\Re$ is not symmetric.

**Example 4.16 (Cont)** It is easy to check that there is only one attack between arguments of $\text{Arg}(\mathcal{K}_i)$: $(\{\text{Bel}_i(\text{Inf}_{j,i}\varphi),\ \text{Bel}_i(\neg\text{Bel}_j\varphi)\},\ \text{Bel}_i(\text{Inf}_{j,i}\varphi \wedge \neg\text{Bel}_j\varphi))\ \Re$ $(\{\text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi)\},\ \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi)$. Thus, $\Re$ is not symmetric.

Next we show that the relation $\Re$ may admit self-attacking arguments.

Example 4.18 Let us consider the following base:

$$\mathcal{K}_i = \begin{cases} \text{TrustSinc}(i, j, \phi) \\ \text{Bel}_i((\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi) \rightarrow \text{Bel}_i(\neg\text{Bel}_j\varphi)) \\ \text{Bel}_i(\text{Inf}_{j,i}\varphi) \end{cases}$$

The argument $(\{\text{TrustSinc}(i, j, \phi),\ \text{Bel}_i(\text{Inf}_{j,i}\varphi),\ \text{Bel}_i((\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi) \rightarrow \text{Bel}_i(\neg\text{Bel}_j\varphi))\},\ \text{Bel}_i(\text{Inf}_{j,i}\varphi \wedge \neg\text{Bel}_j\varphi))$ sinc-attacks itself.

An argumentation system for reasoning about the beliefs of an agent is defined as follows.

Definition 4.19 An *argumentation system* built over a belief base $\mathcal{K}_i$ is a pair $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ where $\Re \subseteq \text{Arg}(\mathcal{K}_i) \times \text{Arg}(\mathcal{K}_i)$ is as given in Definition 4.17.

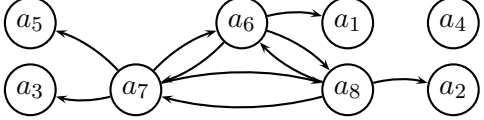
Since arguments may be conflicting, it is important to define the acceptable ones. For that purpose, we use the stable semantics proposed in Dung (1995). This semantics allows to partition the powerset of the set of arguments into two sets: stable extensions and non-extensions. The extensions are used in order to define the inferences to be drawn from the belief base $\mathcal{K}_i$ of agent $i$. These inferences represent what agent $i$ *should believe* according to the available information. The idea is that a formula is inferred if it is supported by at least one argument in every extension. Note that the argument *needs not to be the same in all the extensions.*

Definition 4.20 Let $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$ and $\text{Ext}(\mathcal{T})$ its set of stable extensions. A formula $\phi \in \mathcal{L}$ is *inferred* from $\mathcal{K}_i$ iff for all $\mathcal{E} \in \text{Ext}(\mathcal{T})$, there exists $(H, \phi) \in \mathcal{E}$.
$\text{Output}(\mathcal{T})$ denotes the set of all beliefs inferred from $\mathcal{K}_i$ using system $\mathcal{T}$.

**Example 4.9 (Cont)** Let us consider the belief base $\mathcal{K}_i$ of agent $i$. The set $\text{Arg}(\mathcal{K}_i)$ of arguments is infinite. It contains among others the following arguments:

| |
|---|
| $a_1 : (\{\text{Bel}_i\neg\phi\}, \text{Bel}_i\neg\phi)$ |
| $a_2 : (\{\text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \phi)\}, \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \phi))$ |
| $a_3 : (\{\text{Bel}_i(\text{Inf}_{j,i}\phi)\}, \text{Bel}_i(\text{Inf}_{j,i}\phi))$ |
| $a_4 : (\{\text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi)\}, \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi))$ |
| $a_5 : (\{\text{Bel}_i(\text{Inf}_{j,i}\phi), \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi)\}, \text{Bel}_i(\text{Bel}_j\phi))$ |
| $a_6 : (\{\text{Bel}_i(\text{Inf}_{j,i}\phi), \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \phi)\}, \text{Bel}_i\phi)$ |
| $a_7 : (\{\text{Bel}_i\neg\phi, \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \phi)\}, \{\text{Bel}_i(\neg\text{Inf}_{j,i}\phi))$ |
| $a_8 : (\{\text{Bel}_i\neg\phi, \text{Bel}_i(\text{Inf}_{j,i}\phi)\}, \text{Bel}_i\neg(\text{Inf}_{j,i}\phi \rightarrow \phi))$ |

The following figure summarizes the attacks between the eight arguments:




It can be checked that the argumentation system $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ has three stable extensions. Note that we do not provide the complete result since $\text{Arg}(\mathcal{K}_i)$ is infinite, but give some insights on the arguments that are included in the extensions. Below, if an argument $a_i$ $(i = 1 \ldots 8)$ does not appear in an extension, then it does not belong to that extension. For instance, $a_1 \notin \mathcal{E}_1$.

- $\mathcal{E}_1 = \{a_2, a_3, a_4, a_5, a_6, \ldots\}$
- $\mathcal{E}_2 = \{a_1, a_2, a_4, a_7, \ldots\}$
- $\mathcal{E}_3 = \{a_1, a_3, a_4, a_5, a_8, \ldots\}$.

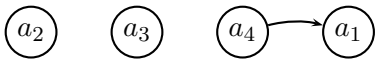
It is worth noticing that the argument $a_4$ belongs to the three extensions. Thus, $\text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi) \in \text{Output}(\mathcal{T})$ meaning that according to the available information, agent $i$ believes in the sincerity of agent $j$ regarding $\phi$. However, $\text{Bel}_i\neg\phi$ and $\text{Bel}_i\phi$ are supported by arguments only in some extensions. Then, $\text{Bel}_i\neg\phi \notin \text{Output}(\mathcal{T})$ and $\text{Bel}_i\phi \notin \text{Output}(\mathcal{T})$ meaning that agent $i$ ignores $\phi$'s truth value.

**Example 4.16 (Cont)**

The table below shows some arguments that may be built from $\mathcal{K}_i$.

| |
|---|
| $a_1 : (\{\text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi)\}, \text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi))$ |
| $a_2 : (\{\text{Bel}_i(\text{Inf}_{j,i}\varphi)\}, \text{Bel}_i(\text{Inf}_{j,i}\varphi))$ |
| $a_3 : (\{\text{Bel}_i(\neg\text{Bel}_j\varphi)\}, \text{Bel}_i(\neg\text{Bel}_j\varphi))$ |
| $a_4 : (\{\text{Bel}_i(\text{Inf}_{j,i}\varphi), \text{Bel}_i(\neg\text{Bel}_j\varphi)\}, \text{Bel}_i(\text{Inf}_{j,i}\varphi \wedge \neg\text{Bel}_j\varphi))$ |

The following figure summarizes the attacks between the four arguments:




It can be checked that the argumentation system $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ has one stable extension: $\mathcal{E} = \{a_2, a_3, a_4, \ldots\}$. Thus, $\text{Bel}_i(\text{Inf}_{j,i}\varphi) \in \text{Output}(\mathcal{T})$, $\text{Bel}_i(\neg\text{Bel}_j\varphi) \in \text{Output}(\mathcal{T})$ but $\text{Bel}_i(\text{Inf}_{j,i}\phi \rightarrow \text{Bel}_j\phi) \notin \text{Output}(\mathcal{T})$. This means that agent $i$ will no longer believe in the sincerity of agent $j$ about $\phi$.

### 4.3. *Properties of the system*

Remember that a belief base of an agent may be inconsistent. We show that the set of inferences drawn from that base using the argumentation system is consistent. Before giving the formal result, we start by another property which shows that every stable extension of the system supports a consistent set of beliefs. Note that this property corresponds exactly to the rationality postulate on consistency that was proposed in Caminada and Amgoud (2007) for rule-based logics and generalized later in Amgoud (2013) for Tarskian logics.

Proposition 4.21 Let $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$ and $\text{Ext}(\mathcal{T})$ its set of stable extensions. For all $\mathcal{E} \in \text{Ext}(\mathcal{T})$, the following properties hold:

- The set $\bigcup_{(H_k,h_k)\in\mathcal{E}} H_k$ is consistent.
- The set $\{h \mid \exists(H,h) \in \mathcal{E}\}$ is consistent.

**Proof:** Let $\mathcal{E}$ be a stable extension of $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$. Assume that the set $\bigcup_{(H_k,h_k)\in\mathcal{E}} H_k$ is inconsistent. Thus, $\exists X \subseteq \bigcup_{(H_k,h_k)\in\mathcal{E}} H_k$ such that $X$ is a minimal (wrt set inclusion) inconsistent set. Since each $H_k$ is consistent, then $|X| > 1$. Thus, for all $\text{Bel}(x) \in X$, $X \setminus \{\text{Bel}(x)\}$ is a minimal set such that $X \setminus \{\text{Bel}(x)\} \vdash \text{Bel}(\neg x)$. Then, $(X \setminus \{\text{Bel}(x)\}, \text{Bel}(\neg x))$ and $(\{\text{Bel}(x)\}, \text{Bel}(x))$ are both arguments. Moreover, $(X \setminus \{\text{Bel}(x)\}, \text{Bel}(\neg x))$ assumption-attacks $(\{\text{Bel}(x)\}, \text{Bel}(x))$. Besides, $\exists(H,h) \in \mathcal{E}$ such that $\text{Bel}(x) \in H$. Thus, $(X \setminus \{\text{Bel}(x)\}, \text{Bel}(\neg x))$ assumption-attacks $(H,h)$. Since $\mathcal{E}$ is conflict-free, then $(X \setminus \{\text{Bel}(x)\}, \text{Bel}(\neg x)) \notin \mathcal{E}$ and $\exists(H',h') \in \mathcal{E}$ such that $(H',h')\Re(X \setminus \{\text{Bel}(x)\}, \text{Bel}(\neg x))$. 1) Assume that $(H',h')$ assumption-attacks $(X \setminus \{\text{Bel}(x)\}, \text{Bel}(\neg x))$. Thus, $\exists \text{Bel}x' \in X \setminus \{\text{Bel}(x)\}$ such that $H' \vdash \text{Bel}\neg x'$. However, $\text{Bel}x' \in H$" for some $(H", h") \in \mathcal{E}$. Thus, $(H',h')$ assumption-attacks $(H", h")$. This contradicts the fact that $\mathcal{E}$ is conflict-free. 2) Assume now that $(H',h')$ sinc-attacks $(X \setminus \{\text{Bel}(x)\}, \text{Bel}(\neg x))$. Then, $h' = \text{Bel}(\text{Inf}_{i,j,\varphi} \wedge \neg\text{Bel}_j\varphi)$ and $\text{TrustSinc}(i,j,\phi) \in X \setminus \{\text{Bel}(x)$. So, $\exists(H", h") \in \mathcal{E}$ such that $\text{TrustSinc}(i,j,\phi) \in H"$. Thus, $(H',h')$ assumption-attacks $(H", h")$. This contradicts the fact that $\mathcal{E}$ is conflict-free. The same reasoning holds for the remaining forms of attacks. Then, $\bigcup_{(H_k,h_k)\in\mathcal{E}} H_k$ is consistent. From the previous result, it follows that the set $\{h \mid \exists(H,h) \in \mathcal{E}\}$ is consistent as well. □

It is worth mentioning that the set of formulas used in the arguments of a stable extension is a consistent subbase of the beliefs base $\mathcal{K}_i$ but not necessarily maximal for set inclusion. This is mainly due to the six attack relations which are not based on inconsistency. Example 4.16 shows a case of a system built over a consistent beliefs base. The system has one stable extension $\mathcal{E}$, and it can be checked that its corresponding base, i.e., $\bigcup_{(H_k,h_k)\in\mathcal{E}} H_k$, is different from $\mathcal{K}_i$.

From this property of the system, it follows that the set $\text{Output}(\mathcal{T})$ is also consistent.

Proposition 4.22 Let $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$. The set $\text{Output}(\mathcal{T})$ is consistent.

**Proof:** From Definition 4.20, it follows that $\text{Output}(\mathcal{T}) \subseteq \{h \mid \exists(H,h) \in \mathcal{E}\}$ for any $\mathcal{E} \in \text{Ext}(\mathcal{T})$. Since $\{h \mid \exists(H,h) \in \mathcal{E}\}$ is consistent then so is $\text{Output}(\mathcal{T})$. □

The next property concerns another rationality postulate in Amgoud (2013) which claims that the extensions should be closed *under sub-arguments*. The idea is that accepting an argument in a given extension implies accepting all its sub-parts in that extension.

Proposition 4.23 Let $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$. For all $\mathcal{E} \in \text{Ext}(\mathcal{T})$, if $(H,h) \in \mathcal{E}$ then for all $(H',h') \in \text{Arg}(\mathcal{K}_i)$ such that $H' \subseteq H$, it holds that $(H',h') \in \mathcal{E}$.

**Proof:** Let $\mathcal{E}$ be a stable extension of $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$. Let $(H,h) \in \mathcal{E}$ and $(H',h') \in \text{Arg}(\mathcal{K}_i)$ such that $H' \subseteq H$ and $(H',h') \notin \mathcal{E}$. Then, $\exists(H", h") \in \mathcal{E}$

such that $(H", h")\Re(H', h')$. 1) Assume that $(H", h")$ assumption-attacks $(H', h')$. Then, $\exists \text{Bel}x \in H'$ such that $h" = \text{Bel}\neg x$. But $\text{Bel}x \in H$ since $H' \subseteq H$. So $(H", h")$ assumption-attacks $(H, h)$. This contradicts the fact that $\mathcal{E}$ is conflict-free. 2) Assume now that $(H", h")$ sinc-attacks $(H', h')$. Then, $h" = \text{Bel}(\text{Inf}_{i,j,\varphi} \wedge \neg\text{Bel}_j\varphi)$ and $\text{TrustSinc}(i, j, \phi) \in H'$. Then $\text{TrustSinc}(i, j, \phi) \in H$. Consequently, $(H", h")$ sinc-attacks $(H, h)$. This contradicts the fact that $\mathcal{E}$ is conflict-free. The same reasoning holds for the remaining forms of attacks. □

The next property concerns the third rationality postulate in Amgoud (2013) which claims that the extensions should be closed under the consequence operator, $\vdash$ in our case. This property guarantees that the system does not forget intuitive conclusions. Before presenting the formal result, let us first introduce a useful notation.

**Notation:** For $X \subseteq \mathcal{L}$, $\text{CN}(X) = \{\phi \in \mathcal{L} \mid X \vdash \phi\}$.

Proposition 4.24 Let $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$ and $\text{Ext}(\mathcal{T})$ its set of stable extensions. For all $\mathcal{E} \in \text{Ext}(\mathcal{T})$, $\{h \mid \exists(H, h) \in \mathcal{E}\} = \text{CN}(\{h \mid \exists(H, h) \in \mathcal{E}\})$.

**Proof:** Let $\mathcal{E}$ be a stable extension of the system $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$. Let $X = \{h \mid \exists(H, h) \in \mathcal{E}\}$. Assume that $X \neq \text{CN}(X)$. Thus, $\exists h \in \text{CN}(X)$ and $h \notin X$. Besides, $X \subseteq \bigcup_{(H_k,h_k)\in\mathcal{E}} \text{CN}(H_k) \subseteq \text{CN}(\bigcup_{(H_k,h_k)\in\mathcal{E}} H_k)$. It follows also that $\text{CN}(X) \subseteq \text{CN}(\bigcup_{(H_k,h_k)\in\mathcal{E}} H_k)$ and thus
$h \in \text{CN}(\bigcup_{(H_k,h_k)\in\mathcal{E}} H_k)$. Two possible cases:

1) $h \in \text{CN}(\emptyset)$, $(\emptyset, h) \in \text{Arg}(\mathcal{K}_i)$ but $(\emptyset, h) \notin \mathcal{E}$. This means that $\exists(H', h')\Re(\emptyset, h)$. But the seven attack relations ensure $h' \in \emptyset$ or $h' = \text{Bel}x \in \emptyset$ and $h = \text{Bel}x$. This is impossible.

2) $h \notin \text{CN}(\emptyset)$ and $\exists S \subseteq \bigcup_{(H_k,h_k)\in\mathcal{E}} H_k$ such that $(S, h) \in \text{Arg}(\mathcal{K}_i)$ since $\bigcup_{(H_k,h_k)\in\mathcal{E}} H_k$ is consistent (see Proposition 4.21). Moreover, $(S, h) \notin \mathcal{E}$. Hence, $\exists(H', h') \in \mathcal{E}$ such that $(H', h')\Re(S, h)$. Assume that $\Re$ is assumption attack. Then, $h' = \text{Bel}\neg x \in S$. But, this implies that $\exists(H'', h'') \in \mathcal{E}$ such that $\text{Bel}\neg x \in H''$ meaning that $(H', h')\Re(H'', h'')$. This contradicts the fact that $\mathcal{E}$ is conflict-free. The same reasoning applies for the six remaining relations since they are all based on attacking the support. □

We show next that the set $\text{Output}(\mathcal{T})$ is closed under $\vdash$.

Proposition 4.25 Let $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$ such that $\text{Ext}(\mathcal{T}) \neq \emptyset$. It holds that $\text{Output}(\mathcal{T}) = \text{CN}(\text{Output}(\mathcal{T}))$.

**Proof:** Let $\mathcal{T} = (\text{Arg}(\mathcal{K}_i), \Re)$ be a system built over a beliefs base $\mathcal{K}_i$ such that $\text{Ext}(\mathcal{T}) \neq \emptyset$. It is clear that $\text{Output}(\mathcal{T}) \subseteq \text{CN}(\text{Output}(\mathcal{T}))$.

Assume now that $h \in \text{CN}(\text{Output}(\mathcal{T}))$ and $h \notin \text{Output}(\mathcal{T})$. Then, $\exists h_1, \ldots, h_n \in \text{Output}(\mathcal{T})$ such that $h \in \text{CN}(\{h_1, \ldots, h_n\})$. Besides, $h_1, \ldots, h_n \in \bigcap_{\mathcal{E}_k\in\text{Ext}(\mathcal{T})} \{\phi \mid \exists(H, \phi) \in \mathcal{E}_k\}$. From monotonicity of CN, it follows that: $\text{CN}(\{h_1, \ldots, h_n\}) \subseteq \text{CN}(\bigcap_{\mathcal{E}_k\in\text{Ext}(\mathcal{T})} \{\phi \mid \exists(H, \phi) \in \mathcal{E}_k\})$. It holds also that $h \in \text{CN}(\{\phi \mid \exists(H, \phi) \in \mathcal{E}_1\}) \cap \ldots \cap \text{CN}(\{\phi \mid \exists(H, \phi) \in \mathcal{E}_n\})$. From Proposition 4.24, $h \in \{\phi \mid \exists(H, \phi) \in \mathcal{E}_1\} \cap \ldots \cap \{\phi \mid \exists(H, \phi) \in \mathcal{E}_n\}$. Consequently, $h \in \text{Output}(\mathcal{T})$. □

This means, for instance, that if $\text{TrustSinc}(i, j, \phi) \in \text{Output}(\mathcal{T})$ and

$\text{Bel}_i(\text{Inf}_{j,i}\phi) \in \text{Output}(\mathcal{T})$, then $\text{Bel}_i(\text{Bel}_j\phi) \in \text{Output}(\mathcal{T})$.

## 5. Graded trust in information sources

In most situations it is an over simplification to say that an agent $i$ trusts (or does not trust) another agent $j$. Rather, in informal terms, we may say that $i$ has a limited trust in $j$, or $i$'s trust in $j$ is high. We are thus faced with the question: "*what is the meaning of graded trust?*".

Demolombe (2009) proposed two different answers to this question. The first answer, when trust is represented by a formula of the form $\text{Bel}_i(\phi_j \Rightarrow \psi_j)$, is that $i$ is *uncertain* to be in a world where the set of $\phi_j$ worlds (i.e., the set of worlds where $\phi_j$ is true) is included in the set of $\psi_j$ worlds (the set of worlds where $\psi_j$ is true). For example, agent $i$ may be uncertain about the fact that agent $j$ is sincere about $p$, that is, about the fact that in every circumstance where $j$ informs $i$ about $p$, it is the case that $j$ believes $p$. Here, graded trust can be defined by the strength level of $i$'s belief about $j$'s sincerity. Notice that this uncertainty level refers to $i$'s beliefs and not to the fact that $j$ is more or less sincere. In more formal terms, according to this interpretation, graded trust can be represented by a formula

$$\text{Bel}_i^g(\phi_j \Rightarrow \psi_j)$$

which is read as follows: the *strength level* of $i$'s belief about the fact that "$\phi_j \Rightarrow \psi_j$ is true" is $g$. In the sequel, $\text{Bel}_i^g$ denotes a "*graded belief*" of agent $i$.

The second answer by Demolombe (2009) is: "$i$ believes that the set of $\phi_j$ worlds is partially included in the set of $\psi_j$ worlds". In such a case, the fact that $i$'s trust in $j$'s sincerity is high can be interpreted as: $i$ believes that in almost all circumstances, if $j$ informs $i$ about $p$, then $j$ believes $p$. According to this interpretation trust level refers to the *regularity* level of the relationship between the fact that $\phi_j$ is true and the fact that $\psi_j$ is true. Graded trust is thus formally represented by the formula:

$$\text{Bel}_i(\phi_j \Rightarrow^h \psi_j)$$

where $h$ may be a numerical value which represents *graded regularity*.

For the purpose of our proposal, graded trust may refer to both kinds of levels (uncertainty and regularity). It is thus represented by formulas of the form:

$$\text{Bel}_i^g(\phi_j \Rightarrow^h \psi_j)$$

whose intended meaning is that the strength level of $i$'s belief about the fact that $\phi_j$ entails $\psi_j$ with a regularity level $h$ is $g$. It is worth pointing out that in general these two levels are independent. It may be the case, for example, that $i$ strongly believes that $j$'s sincerity is low or that $i$ strongly believes that $j$'s sincerity is high and it may also be the case that $i$ has a low level of belief about the fact that $j$'s sincerity is low.

### 5.1. *Extended logic*

In what follows, we extend the logical language of Section 2 for reasoning about graded trust. Let us first recall the intuitive meaning of the new operators:

- $\mathrm{Bel}_i^g\phi$: the strength level of $i$'s belief about the fact that $\phi$ is true is (exactly) $g$.
- $\phi \Rightarrow^h \psi$: $\phi$ entails $\psi$ at level $h$.
- $\Box\phi$: $\phi$ holds in all the situations.

The operator $\Box$ is introduced for formal purposes that are explained below. We also assume two additional sets that contain levels of beliefs and regularity:

- $GRB$: finite set of belief levels.
- $GRR$: finite set of regularity levels.

Notice that no particular assumption is made on the nature of the elements of these sets. However, we assume that they are both equipped with a preordering $\leq$ (i.e., a reflexive and transitive binary relation). For $x, y \in GRB$ (respectively in $x, y \in GRR$), $x \leq y$ means that $y$ is at least as strong as $x$. The strict relation associated with $\leq$ is denoted by $<$ and defined as follows: $x < y \stackrel{\text{def}}{=} (x \leq y) \wedge \text{not}(y \leq x)$. Moreover, both sets has a lower and an upper bounds denoted respectively $min$ and $max$[1]. For every $x$ in $GRB$ or in $GRR$, $min \leq x \leq max$.

**Notations:** $Forall(g, cond)F(g) \stackrel{\text{def}}{=} \bigwedge_{g \in G, cond(g)} F(g)$, $Exist(g, cond)F(g) \stackrel{\text{def}}{=} \bigvee_{g \in G, cond(g)F(g)}$, and $\psi^h \stackrel{\text{def}}{=} \top \Rightarrow^h \psi$.

The logic associated with the extended language is based on the following inference rules and axiom schemas.

| | |
|---|---|
| (SubstBel) | If $\vdash \phi \leftrightarrow \psi$ then $\vdash \mathrm{Bel}_i^g(\phi) \leftrightarrow \mathrm{Bel}_i^g(\psi)$ |
| (Weak) | If $\vdash \phi \rightarrow \psi$ then $\vdash \mathrm{Bel}_i^g(\phi) \rightarrow \neg Exist(g', g' < g)\mathrm{Bel}_i^{g'}\psi$ |
| (ClosDisj) | If $g_3 = Max\{g_1, g_2\}$ then $\vdash \mathrm{Bel}_i^{g_1}(\phi_1) \wedge \mathrm{Bel}_i^{g_2}(\phi_2) \rightarrow \mathrm{Bel}_i^{g_3}(\phi_1 \vee \phi_2)$ |
| (ClosConj) | If $g_3 = Min\{g_1, g_2\}$ then $\vdash \mathrm{Bel}_i^{g_1}(\phi_1) \wedge \mathrm{Bel}_i^{g_2}(\phi_2) \rightarrow \mathrm{Bel}_i^{g_3}(\phi_1 \wedge \phi_2)$ |
| (UnicBel) | $\vdash Forall(g_1, g_2, g_1 \neq g_2)\ \neg(\mathrm{Bel}_i^{g_1}(\phi) \wedge \mathrm{Bel}_i^{g_2}(\phi))$ |
| (Consist) | $\vdash \mathrm{Bel}_i^g\phi \rightarrow \mathrm{Bel}_i\phi$ |
| (MinBel) | $\vdash (\mathrm{Bel}_i^{min}\phi \wedge \mathrm{Bel}_i^g\psi) \rightarrow \Box(\phi \rightarrow \psi)$ |
| (MaxBel) | $\vdash (\mathrm{Bel}_i^{max}\phi \wedge \mathrm{Bel}_i^g\psi) \rightarrow \Box(\psi \rightarrow \phi)$ |
| (MaxTau) | If $\vdash \phi$ then $\vdash \mathrm{Bel}_i^{max}\phi$ |
| (PosInt) | $\vdash \mathrm{Bel}_i^g(\phi) \rightarrow \mathrm{Bel}_i\mathrm{Bel}_i^g(\phi)$ |
| (NegInt) | $\vdash \neg\mathrm{Bel}_i^g(\phi) \rightarrow \mathrm{Bel}_i\neg\mathrm{Bel}_i^g(\phi)$ |
| (SubstReg) | If $\vdash \phi \leftrightarrow \phi'$ and $\vdash \psi \leftrightarrow \psi'$ then $\vdash (\phi \Rightarrow^h \psi) \rightarrow (\phi' \Rightarrow^h \psi')$ |
| (Detach) | $\vdash (\phi \Rightarrow^h \psi) \rightarrow (\phi \rightarrow \psi^h)$ |
| (Trans) | If $n = Max\{Min\{h_1, k_1\}, Min\{h_2, k_2\}\}$, then $\vdash ((\phi \Rightarrow^{h_1} \psi) \wedge (\phi \wedge \psi \Rightarrow^{k_1} \theta) \wedge (\phi \Rightarrow^{h_2} \neg\psi) \wedge (\phi \wedge \neg\psi \Rightarrow^{k_2} \theta)) \rightarrow (\phi \Rightarrow^n \theta)$ |
| (UnicReg) | $\vdash Forall(h_1, h_2, h_1 \neq h_2)\ \neg((\phi \Rightarrow^{h_1} \psi) \wedge (\phi \Rightarrow^{h_2} \psi))$ |
| (MinReg) | $\vdash (\phi \Rightarrow^{min} \psi) \leftrightarrow \Box(\phi \rightarrow \neg\psi)$ |
| (MaxReg) | $\vdash (\phi \Rightarrow^{max} \psi) \leftrightarrow \Box(\phi \rightarrow \psi)$[2] |
| (DetachBel) | If $\mathrm{Bel}_i^g\phi$ and $\vdash \phi \rightarrow \psi$, then $\neg Exist(g', g' < g)\mathrm{Bel}_i^{g'}\psi$. |

The first rule, (SubstBel), states that in $\mathrm{Bel}_i^g(\phi)$, $\phi$ can be substituted by any logically equivalent formula. (Weak) says that if $\psi$ is a logical consequence of $\phi$ (i.e. $\vdash \phi \rightarrow \psi$), then, if $i$ has ascribed a strength level to his belief about $\phi$ and to his

[1] We use the same notations for the minimal element and for the maximal element in $GRB$ and in $GRR$ while they are not necessarily identical. The context allows us to avoid ambiguities.

belief about $\psi$, then the level of $\psi$ cannot be lower than the level of $\phi$. (ClosDisj) says that if the levels of beliefs of two formulas $\phi_1$ and $\phi_2$ are fixed, then the level of their disjunction is the maximum of these two levels. With (ClosConj) schema, if the levels of beliefs of two formulas $\phi_1$ and $\phi_2$ are fixed, then the level of their conjunction is the minimum of these two levels. (UnicBel) states that the strength level of $i$'s belief is unique for every sentence. According to (Consist) schema, graded beliefs are considered as standard beliefs to which an agent $i$ has assigned a strength level. It may be that $i$ has not assigned a strength level to some belief, for instance, because he has no argument to assign it such or such level. According to this axiom schema $\mathrm{Bel}_i^g(\phi)$ can be rephrased as: $i$ believes $\phi$ and the strength level of this belief is $g$. The axiom schema (MinBel) states that if $\phi$ represents the formula which is believed at the minimum level and $\psi$ is believed at some belief level, then $\phi$ implies $\psi$. It is worth noticing that this axiom is consistent with (ClosConj). From an intuitive point of view a formula which is believed at the minimal level is a formula which denotes a proposition which is more specific than any other formula which is believed at any other level. That means that the set of $\phi$ worlds is included into the set of $\psi$ worlds. (MaxBel) says that if $\phi$ represents the formula which is believed at the maximum level and $\psi$ is believed at some belief level, then $\psi$ implies $\phi$. This axiom schema is consistent with (ClosDisj). From an intuitive point of view a formula which is believed at the maximal level is a formula which denotes a proposition which is less specific than any other formula which is believed at any other level. This means that the set of $\phi$ worlds contains the set of $\psi$ worlds. The schema (MaxTau) states that if $\phi$ is a theorem of the logic, then the belief level of $\phi$ is $max$. According to schema (PosInt), if a formula $\phi$ is believed at level $g$, then $i$ believes, in the standard sense, that $\phi$ is believed at level $g$. This positive introspection axiom schema means that no level is ascribed by $i$ to his evaluation of the level of a belief. If such a level would be ascribed, one could ask the question: *what is i's evaluation of this "second" order level?*, and we would be led to an infinite number of introspection levels, which is far to be intuitive. (NegInt) says that if formula $\phi$ is not believed at level $g$, then $i$ believes, in the standard sense, that $\phi$ is not believed at level $g$. Note that the justification of (NegInt) is similar to the justification of (PosInt). The axiom (SubstReg) concerns the conditional connective $\Rightarrow^h$; it says that in the formula $\phi \Rightarrow^h \psi$, both $\phi$ and $\psi$ can be substituted by logically equivalent formulas. (Detach) states that if $\phi$ entails $\psi$ at level $h$, then if $\phi$ holds, $\psi$ holds at level $h$. Note that "$\psi$ holds at level $h$" is an abbreviation for "$True \Rightarrow^h \phi$". The axiom (Trans) says that there exists a function $F$ such that if $n = F(h_1, k_1, h_2, k_2)$, then if $\phi$ entails $\psi$ at level $h_1$, $\phi \wedge \psi$ entails $\theta$ at level $k_1$, $\phi$ entails $\neg\psi$ at the level $h_2$ and $\phi \wedge \neg\psi$ entails $\theta$ at level $k_2$, then $\phi$ entails $\theta$ at level $n$. This axiom seems quite complex but it is mandatory since, in general, from $(\phi \Rightarrow^{h_1} \psi) \wedge (\phi \wedge \psi \Rightarrow^{k_1} \theta)$, we cannot infer what is the value of $n$ such that: $\phi \Rightarrow^n \theta$, because there may be $\phi$ worlds that are $\theta$ worlds and which are not $\psi$ worlds. Notice that axiom schema (Trans) is perfectly compatible with conditional probabilities if we accept some uniform distribution assumptions. In this case, the form of $F$ is $n = (h_1 \times k_1) + (h_2 \times k_2)$. Even if this is not a sufficient justification, by analogy with conditional probabilities we have adopted the following function $F$: $n = Max\{Min\{h_1, k_1\},\ Min\{h_2, k_2\}\}$. The axiom (UnicReg) states that the regularity level of $\phi$ entails $\psi$ is unique whereas (MinReg) ensures that $\phi$ entails $\psi$ at the minimum level iff $\phi$ implies $\neg\psi$. The intuitive idea is that $\phi \Rightarrow^{min} \psi$ holds iff the set of $\phi$ worlds and the set of $\psi$ worlds are disjoint. The sentence $\phi \Rightarrow^{min} \psi$ can be interpreted in the context of conditional probabilities as $0 = Pr(\psi|\phi)$. According to axiom (MaxReg), a formula $\phi$ entails $\psi$ at the maximum level iff $\phi$ implies $\psi$. The intuitive idea is

that $\phi \Rightarrow^{max} \psi$ holds iff the set of $\phi$ worlds is included into the set of $\psi$ worlds. Note that the sentence $\phi \Rightarrow^{max} \psi$ can be interpreted in the context of conditional probabilities as $1 = Pr(\psi|\phi)$. The last axiom (DetachBel) follows from the axioms (MaxTau), (ClosConj), (SubstBel) and (Weak).

In the sequel, $\mathcal{L}'$ will denote the extended language and $\vdash^*$ the extended logic, i.e., the logic $\vdash$ extended with the previous axioms.

### 5.2. *Preference-based argumentation for graded trust*

There is a clear consensus in the literature that arguments do not necessarily have the same strength. It may be the case that an argument relies on certain information while another argument is built from less certain ones, or that an argument promotes an important value while another promotes a weaker one. In both cases, the former argument is clearly stronger than the latter. These differences in arguments' strengths make it possible to compare them. Consequently, several preference relations between arguments have been defined in the literature (e.g. Amgoud (1999), Benferhat et al. (1993), Cayrol et al. (1993), Simari and Loui (1992)). There is also a consensus on the fact that preferences should be taken into account in the evaluation of arguments (see Amgoud and Cayrol (2002), Bench-Capon (2003), Modgil (2009), Prakken and Sartor (1997), Simari and Loui (1992)).

In Amgoud and Cayrol (2002), a first *abstract* preference-based argumentation framework was proposed. It takes as input a set of arguments, an attack relation, and a preference relation $\succeq$ between arguments. For two arguments $a$ and $b$, $a \succeq b$ means that the argument $a$ is at least as strong as $b$. The relation $\succeq$ is abstract and can be instantiated in different ways. However, it is assumed to be a (total or partial) pre-ordering (i.e., reflexive and transitive). The strict version associated with $\succeq$ is denoted by $\succ$ and is defined as follows: $a \succ b$ iff $a \succeq b$ and not $b \succeq a$. Whatever the source of this preference relation is, the idea is to ignore an attack if the attacked argument is stronger than its attacker. Dung's semantics are applied on the remaining attacks. This approach is particularly interesting when the attack relation is symmetric. However, when the attack relation is not symmetric like the relation given in Definition 4.17, the extensions of the argumentation framework may be conflicting leading thus to counter-intuitive results. Consequently, Amgoud and Vesic (2009) proposed a new approach which consists of inverting the direction of an attack whenever the attacker is weaker than its target as follows:

Definition 5.1 Let $(\mathcal{A}, \mathcal{R}, \succeq)$ be an argumentation framework. For two arguments $a, b \in \mathcal{A}$, $a$ defeats $b$ iff

- $a\mathcal{R}b$ and not $b \succ a$ or
- $b\mathcal{R}a$ and $a \succ b$

Dung's semantics are then applied to the new framework $(\mathcal{A}, \text{defeats})$ for evaluating the arguments. In what follows, we propose an instantiation of this abstract framework for reasoning about graded trust. As in the binary case, we assume a knowledge base $\mathcal{K}_i$ containing the beliefs of an agent $i$. Formulas of $\mathcal{K}_i$ are elements of the extended language $\mathcal{L}'$. Arguments are built from $\mathcal{K}_i$ following Definition 4.4, however by replacing the relation $\vdash$ by $\vdash^*$.

Definition 5.2 Let $\mathcal{K}_i$ be a beliefs base of agent $i$. An *argument* is a pair $(H, h)$

where:

- $H \subseteq \mathcal{K}_i$ and $h \in \mathcal{L}'$
- $H$ is consistent
- $H \vdash^* h$
- $\nexists H' \subset H$ such that $H' \vdash^* h$

Arguments attack each other as shown in Definition 4.17, i.e., the relation used in the binary case. However, they may have different strength levels. It is the strength level of the weakest (or the less certain) formula used in its support.

Definition 5.3 Let $(H, h)$ be an argument such that $H = \{\text{Bel}_i^{g_1}\phi_1, \ldots, \text{Bel}_i^{g_n}\phi_n\}$. The *strength level* of $(H, h)$, denoted $\text{Level}(H, h)$, is $\text{Min}\{g_1, \ldots, g_n\}$.

These strengths are used in order to compare arguments. The idea is to prefer the one with the greatest strength level, i.e., the one whose support is based on more certain information.

Definition 5.4 Let $(H, h), (H', h')$ be two arguments. $(H, h) \succeq (H', h')$ iff $\text{Level}(H, h) \geq \text{Level}(H', h')$.

The argumentation framework $(\text{Arg}(\mathcal{K}_i), \Re, \succeq)$ is used for reasoning about the beliefs base $\mathcal{K}_i$ of agent $i$. Let us illustrate this framework on a simple example.

Example 5.5 Let us consider the following base:

$$\mathcal{K}_i = \begin{cases} Bel_i^{g_4}(parking) \\ Bel_i^{g_2}(parking \rightarrow office) \\ Bel_i^{g_3}(meeting) \\ Bel_i^{g_4}(meeting \rightarrow \neg office) \end{cases}$$

where it is assumed that the set of strengths of beliefs is $\{g_1, g_2, g_3, g_4, g_5\}$ and this set has the structure of a total order. Moreover, the following abbreviations have been adopted: *parking*: Luis's car is at the parking, $office$: Luis is at his office, *meeting*: Luis is attending a meeting, and *teaching*: Luis is teaching.

The intuitive justifications of the beliefs strength levels are that agent $i$ has observed that Luis's car is at the parking and $i$ is not strongly convinced that this fact guarantees that Luis is at his office. Moreover, $i$ has been informed that Luis is attending a meeting and $i$ knows that a meeting cannot happens at Luis's office.

From $\mathcal{K}_i$, an infinite number of arguments can be built including the following ones:

$a_1 = (H_1, h_1)$, where $H_1$ is $\{Bel_i^{g_4}(parking), Bel_i^{g_2}(parking \rightarrow office)\}$ and $h_1 = Bel_i^{g_2}(parking \wedge office)$.

From $Level(a)$ definition we have: $Level(a_1) = min\{g_2, g_4\} = g_2$.

From the inference rules (SubstBel) and (Weak), we also have the argument:

$a_2 = (H_1, h_2)$, where $h_2 = Bel_i^{g}(office)$ and $g$ is greater or equal to $g_2$. We also have: $Level(a_2) = g_2$. Notice that the strength of the consequence $h_2$ is not necessarily the same as the strength of its support $H_1$.

We also have the arguments:

$a_3 = (H_3, h_3)$, where $H_3$ is $\{Bel_i^{g_3}(meeting), Bel_i^{g_4}(meeting \rightarrow \neg office)\}$ and $h_3 = Bel_i^{g_3}(meeting \wedge \neg office)$. We have $Level(a_3) = g_3$.

$a_4 = (H_3, h_4)$, where $h_4 = Bel_i^{g'}(\neg office)$ and $g'$ is greater or equal to $g_3$.

In the logic presented in section 5 graded beliefs are assumed to be standard beliefs (see schema (Consist)) and standard beliefs must be consistent in the sense of schema (D). According to this logic arguments $a_2$ and $a_4$ lead to an inconsistency. This kind of inconsistency can be removed if it is accepted that $Bel_i^g\phi$ means that the strength level of the fact $i$ believes that $\phi$ may be true is $g$ (instead of $i$ believes that $\phi$ is true is $g$). This interpretation of $Bel_i^g\phi$ can be formally represented replacing the schema (Consist) by: (Consist') $\vdash Bel_i^g\phi \rightarrow \neg Bel_i\neg\phi$.

In the same context we could have the following knowledge base $\mathcal{K}_i'$ where agent $i$ trusts agent $j$ in his validity about *meeting* and $j$ has informed $i$ about *meeting*.

$\mathcal{K}_i' = \{Bel_i^{g_4}(parking), Bel_i^{g_2}(parking \rightarrow office), Bel_i^{g_4}(Inf_{j,i}meeting),$
$Bel_i^{g_2}(Inf_{j,i}meeting \rightarrow meeting), Bel_i^{g_4}(meeting \rightarrow \neg office)\}$

In $\mathcal{K}_i'$ we have the argument $a_5$.

$a_5 = (H_5, h_5)$, where $H_5$ is $\{Bel_i^{g_4}(Inf_{j,i}meeting), Bel_i^{g_2}(Inf_{j,i}meeting \rightarrow meeting), Bel_i^{g_4}(meeting \rightarrow \neg office)\}$ and $h_5 = Bel_i^{g_2}(Inf_{j,i}meeting \wedge meeting \wedge \neg office)$ and $Level(a_5) = g_2$.

We may have a more complex knowledge base $\mathcal{K}_i''$ where is represented the fact that if Luis is teaching, he cannot be attending a meeting.

$\mathcal{K}_i'' = \{Bel_i^{g_4}(parking), Bel_i^{g_2}(parking \rightarrow office), Bel_i^{g_4}(Inf_{j,i}meeting),$
$Bel_i^{g_2}(Inf_{j,i}meeting \rightarrow meeting), Bel_i^{g_4}(meeting \rightarrow \neg office), Bel_i^{g_4}(teaching), Bel_i^{g_4}(teaching \rightarrow \neg meeting)\}$

Now, we have the argument $a_6$:

$a_6 = (H_6, h_6)$ where $H_6$ is $\{Bel_i^{g_4}(teaching), Bel_i^{g_4}(teaching \rightarrow \neg meeting)\}$ and $h_6 = Bel_i^{g_4}(teaching \wedge \neg meeting)$ and $Level(a_6) = g_4$.

Since the consequence $h_6$ of $a_6$ is $Bel^{g_4}(teaching \wedge \neg meeting)$ and in the support $H_3$ of $a_3$ we have $Bel_i^{g_3}(meeting)$, we can accept, thanks to a limited change in the attack definition, that $a_6$ attacks $a_3$. Moreover, we have $Level(a_6) > Level(a_3)$, then we can infer that $a_6$ defeats $a_3$.

## 6. Related work

Trust modeling has become a hot topic during the last ten years. More than twenty definitions were proposed for this complex concept. Among others the following one was proposed by Falcone and Castelfranchi (2001):

> Trust is a mental state, a complex attitude of an agent $i$ towards another agent $j$ about the behavior/action $a$ relevant for the goal $g$.

Gambetta (1990) defines trust as a subjective probability by which an agent $i$ expects that another agent $j$ performs a given action on which its welfare depends. In Liau (2003), trust is represented as agent's beliefs and the author focused on trust in validity and its impact on the assimilation of information received from the trustee. The basic idea is the following: if agent $i$ believes that agent $j$ has told him the truth of $\phi$ and $i$ trusts the judgment of $j$ on $\phi$, then he will also believe $\phi$. Our formalism follows this line of research and considers six forms of trust including validity, sincerity, and competence. It shows how to build arguments in favor (respectively against) each form of trust, and how to use beliefs concerning

the trustworthiness of the other agents in order to infer new beliefs.

Some attempts on combining argumentation theory and trust have been made in the literature. Based on the representation proposed in Liau (2003), Villata et al. (2011) presented an instantiation of the meta-argumentation model (Boella et al. (2009)) for reasoning about trust in validity. The technique of meta-argumentation applies Dung's theory of abstract argumentation to itself. The instantiation contains arguments built from beliefs and *meta-arguments*. An example of a meta argument is of the form Trust $i$ meaning that "agent $i$ is trustable". Our formalism is more general since it reasons about more forms of trust. Moreover, it is much more simple since it instantiates directly Dung's framework with a clear and intuitive logical language in which various kinds of beliefs are represented.

An argumentation-based model for reasoning about inconsistent and uncertain information was proposed in Tang et al. (2012). It is as an instantiation of the preference-based argumentation framework proposed in Amgoud and Cayrol (2002) where arguments do not necessarily have the same strengths and are thus compared using a binary relation expressing *preferences*. The arguments are built from a base which contains beliefs pervaded with degrees of certainty. These degrees are then combined for computing the certainty levels of the supports of arguments which in turn are used for comparing arguments. The particularity of the model is the use of trusted information in order to assign degrees for inferred beliefs. Indeed, the model takes as input a simple network whose nodes are agents and edges represent trust relationships between nodes. For instance, an arc from agent $i$ towards agent $j$ means that agent $i$ trusts agent $j$. Weights are associated with edges and express degrees of trust. Our formalism is based on a richer model of trust. It distinguishes between six forms of trusts instead of an absolute trust in Tang et al. (2012). Moreover, our formalism not only uses trusted information in order to infer new beliefs but also reasons about trust itself and infers beliefs about trust.

More recently, in Parsons et al. (2012) the authors focused on identifying ten sources of trust and presented them in terms of *argument schemes*, i.e., syllogisms justifying trustworthiness in an agent. Examples of sources are authority, reputation and expert opinion which is called in our formalism competence. Critical questions showing how each argument scheme can be attacked were also proposed. While some of the proposed sources make sense, others are debatable. For instance, trust because of *pragmatism* says that an agent $i$ may decide to trust another agent $j$ because it serves $i$'s interests to do so. There is a form of wishful thinking which is not compatible with the fact that trust is a belief.

Another interesting contribution on the combination of argumentation theory and trust was done in Stranders et al. (2007). The focus is on computing to what extent agent $i$ trusts agent $j$. This is done from statistical data and arguments. The model is an instantiation of the abstract decision model proposed in Amgoud and Prade (2009). Our formalism does not use statistical data. Moreover, it is an inference model and not a decision making one.

Finally, in Matt et al. (2010) the authors proposed a model for evaluating the trust an agent may have in another. For that purpose, arguments in favor of trust are built. They are mainly grounded on statistical data which makes this approach different from the one we followed in the present paper.

## 7. Conclusion

This paper tackled the important questions of formalizing and reasoning about trust in information sources. It proposed a formal model based on the construction and evaluation of arguments. The model presents several advantages: first, it is grounded on an accurate and simple logical language for representing trust in information sources. Indeed, modal logic is used for distinguishing between what is true (respectively false) and what is believed by an agent. Second, unlike existing works that define absolute trust in an agent, our model uses a fine-grained notion of trust. It distinguishes between six forms of trust including trust in the sincerity of an agent and trust in his competence. The third feature of our model is that it plays two distinct roles: i) it shows how to take into account trust in information sources in order to deal and reason about information coming from those sources, ii) it shows whether to trust or not a given source of information on the basis of available beliefs. This makes our model a good candidate for dialog systems.

There are a number of ways to extend this work. Our future direction consists of investigating the properties of the model under other semantics, namely preferred semantics. We have shown that the attack relations we have defined are very special since they are not grounded on inconsistency. Consequently, despite the fact that arguments are consistent, self-attacking arguments may exist preventing thus the existence of stable extensions.

Another interesting future direction consists of refining the logical language by considering the notion of *topic*. The basic idea is to represent information such as: Agent $i$ trusts the competence of agent $j$ in psychology but not in philosophy. Our formal definitions can be extended in this direction thanks to the logic of *aboutness* developed by Demolombe and Jones (1995). The logical language of this logic contains a predicate $A(t, \phi)$ whose intuitive meaning is that formula $\phi$ is about topic $t$. This predicate can be used, for instance, for expressing the fact that $i$ trusts $j$ in his validity for any sentence about a given topic $t$: $\forall x(A(t, x) \rightarrow \text{TrustVal}(i, j, x)$. Another direction consists of handling *graded* trust. In the proposed model, trust is a binary notion: an agent either fully trusts another agent or fully distrust the agent. However, in everyday life one may have a limited trust in a person. It is thus important to define to what extent an agent trusts another.

## References


Amgoud, L. (1999), "Contribution à l'intégration des préférences dans le raisonnement argumentatif," *Thèse de doctorat, Université Paul Sabatier, Toulouse, France.*

Amgoud, L. (2013), "Postulates for logic-based argumentation systems," *Journal of Approximate Reasoning.*

Amgoud, L., and Besnard, P. (2009), "Bridging the gap between abstract argumentation systems and logic," in *International Conference on Scalable Uncertainty Management, SUM'09*, pp. 12–27.

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

Amgoud, L., Maudet, N., and Parsons, S. (2000), "Modelling dialogues using argumentation," in *Proceedings of the 4th International Conference on MultiAgent Systems (ICMAS'00), IEEE*, pp. 31–38.

Amgoud, L., and Prade, H. (2009), "Using arguments for making and explaining decisions," *Artificial Intelligence Journal*, 173, 413–436.

Amgoud, L., and Vesic, S. (2009), "Repairing Preference-Based Argumentation Systems," in *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'09)*, pp. 665–670.

Baroni, P., Giacomin, M., and Guida, G. (2005), "SCC-recursiveness: a general schema for argumentation semantics," *Artificial Intelligence Journal*, 168, 162–210.

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

Benferhat, S., Dubois, D., and Prade, H. (1993), "Argumentative inference in uncertain and inconsistent knowledge bases," in *Proceedings of the 9th Conference on Uncertainty in Artificial intelligence*

*(UAI'93)*, pp. 411–419.
Black, E., and Hunter, A. (2009), "An inquiry dialogue system," *Autonomous Agents and Multi-Agent Systems*, 19, 173–209.
Boella, G., Gabbay, D., van der Torre, L., and Villata, S. (2009), "Meta-Argumentation Modelling I: Methodology and Techniques," *Studia Logica*, 93, 297–355.
Caminada, M., and Amgoud, L. (2007), "On the evaluation of argumentation formalisms," *Artificial Intelligence Journal*, 171 (5-6), 286–310.
Castelfranchi, C. (2011), "Trust: nature and dynamics," in *ACM SIGCHI Italian Chapter International Conference on Computer-Human Interaction*, pp. 13–14.
Castelfranchi, C., and Falcone, R. (2000), "Trust is Much More Than Subjective Probability: Mental Components and Sources of Trust," in *HICSS*.
Cayrol, C., Royer, V., and Saurel, C. (1993), "Management of preferences in Assumption-Based Reasoning," *Lecture Notes in Computer Science*, 682, 13–22.
Chellas, B., *Modal logic: an introduction*, Cambridge University Press, Cambridge (1980).
Demolombe, R. (1998), "To trust information sources: a proposal for a modal logical framework," in *Proc. of the Workshop on Deception, Fraud and Trust in Agent Societies*, ed. C. Castelfranchi and Y-H. Tan.
Demolombe, R. (1999), "To trust information sources: a proposal for a modal logical framework," in *Trust and Deception in Virtual Societies, Kluwer, Dordrecht*.
Demolombe, R. (2004), "Reasoning About Trust: A Formal Logical Framework," in *Second International Conference on Trust Management, iTrust'04*, pp. 291–303.
Demolombe, R. (2009), "Graded Trust," in *Proceedings of the Trust in Agent Societies Workshop at AAMAS 2009*, ed. R. Falcone and S. Barber and J. Sabater-Mir and M.Singh.
Demolombe, R. (2011), "Transitivity and Propagation of Trust in Information Sources: An Analysis in Modal Logic," in *Workshop on Computational Logic in Multi-Agent Systems*, pp. 13–28.
Demolombe, R., and Jones, A. (1995), "Reasoning about Topics: towards a formal theory," in *American Association for Artificial Intelligence Fall Symposium*.
Demolombe, R., and Liau, C.J. (2001), "A logic of graded trust and belief fusion," in *Proceedings of 4th Workshop on Deception, Fraud and Trust*, eds. C. Castelfranci and R. Falcone.
Demolombe, R., and Lorini, E. (2008), "A logical account of trust in information sources," in *International Workshop on Trust in Agent Societies*.
Dung, P.M. (1995), "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and $n$-person games," *Artificial Intelligence Journal*, 77, 321–357.
Dung, P., Mancarella, P., and Toni, F. (2007), "Computing ideal skeptical argumentation," *Artificial Intelligence Journal*, 171, 642–674.
Elvang-Gøransson, M., Fox, J., and Krause, P. (1993), "Acceptability of arguments as 'logical uncertainty," in *Proceedings of the 2nd European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'93*, pp. 85–90.
Falcone, R., and Castelfranchi, C. (2001), "Social trust: a cognitive approach," *Trust and deception in virtual societies*, pp. 55–90.
Falcone, R., Piunti, M., Venanzi, M., and Castelfranchi, C. (2013), "From manifesta to krypta: The relevance of categories for trusting others," *ACM TIST*, 4, 27.
Gambetta, D. (1990), "Can we trust them?," *Trust: Making and breaking cooperative relations*, pp. 213–238.
Gorogiannis, N., and Hunter, A. (2011), "Instantiating abstract argumentation with classical logic arguments: Postulates and properties," *Artificial Intelligence Journal*, 175 (9–10), 1479–1497.
Jennings, N.R., Mamdani, E.H., Corera, J., Laresgoiti, I., Perriolat, F., Skarek, P., and Varga, L.Z. (1996), "Using ARCHON to develop real-word DAI applications Part 1," *IEEE Expert*, 11, 64–70.
Liau, C. (2003), "Belief, information acquisition, and trust in multi-agent systems–A modal logic formulation," *Artificial Intelligence Journal*, 149, 31–60.
Lorini, E., and Demolombe, R. (2008), "From Binary Trust to Graded Trust in Information Sources: A Logical Perspective," in *11th International Workshop on Trust in Agent Societies*, pp. 205–225.
Maes, P. (1996), "Agents that reduce work and information overload," *Communication of the ACM*, 37(7), 31–40.
Marsh, S. (1994), "Formalising trust as a computational concept," Technical report, Ph.D. Thesis, University of Stirling.
Matt, P., Morge, M., and Toni, F. (2010), "Combining statistics and arguments to compute trust," in *AAMAS*, pp. 209–216.
McBurney, P., Hitchcock, D., and Parsons, S. (2007), "The eightfold way of deliberation dialogue," *International Journal of Intelligent Systems*, 22, 95–132.
Modgil, S. (2009), "Reasoning about preferences in argumentation frameworks," *Artificial Intelligence Journal*, 173:9–10, 901–934.
Parsons, S., Atkinson, K., Haigh, K., Levitt, K., McBurney, P., Rowe, J., Singh, M., and Sklar, E. (2012), "Argument Schemes for Reasoning about Trust," in *Computational Models of Argument, COMMA'12*, pp. 430–441.
Prakken, H., and Sartor, G. (1997), "Argument-based extended logic programming with defeasible priorities," *Journal of Applied Non-Classical Logics*, 7, 25–75.
Rodriguez, J.A., Noriega, P., Sierra, C., and Padget, J. (1997), "A Java-based electronic auction house," in *Proceedings of the 2nd International Conference on the Practical Application of intelligent Agents and Multi-Agent Technology*, pp. 207–224.
Shi, J., Bochmann, G., and Adams, C. (2005), "A Trust Model with Statistical Foundation," *IFIP Advances in Information and Communication Technology*, 173, 145–158.
Simari, G., and Loui, R. (1992), "A mathematical treatment of defeasible reasoning and its implementation," *Artificial Intelligence Journal*, 53, 125–157.
Stranders, R., de Weerdt, M., and Witteveen, C. (2007), "Fuzzy Argumentation for Trust," in *International*

*Workshop on Computational Logic in Multi-Agent Systems, CLIMA'07*, pp. 214–230.

Sycara, K. (1990), "Persuasive argumentation in negotiation," *Theory and Decision*, 28, 203–242.

Tang, Y., Cai, K., McBurney, P., Sklar, E., and Parsons, S. (2012), "Using argumentation to reason about trust and belief," *Journal of Logic and Computation*, 22, 979–1018.

Tarski, A. (1956), "On Some Fundamental Concepts of Metamathematics," *Logic, Semantics, Metamathematics (E. H. Woodger, editor))*, Oxford Uni. Press.

Villata, S., Boella, G., Gabbay, D., and van der Torre, L. (2011), "Arguing about the Trustworthiness of the Information Sources," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'11*, pp. 74–85.

Walton, D.N., and Krabbe, E.C.W., *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*, Albany, NY: State University of New York Press (1995).

Wellman, M.P. (1993), "A market-oriented programming environment and its application to distributed multicommodity flow problems," *Artificial Intelligence and Research*, 1, 1–23.

Wooldridge, M.J., and Jennings, N. (1995), "Intelligent agents: theory and practice," *The Knowledge Engineering Review*, 10, 115–152.