Arguing about Trust in Information Sources

Leila Amgoud
IRIT - CNRS
118, route de Narbonne
31062, Toulouse-France
amgoud@irit.fr

Robert Demolombe
IRIT - CNRS
118, route de Narbonne
31062, Toulouse-France
robert.demolombe@orange.fr

ABSTRACT

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 the *trustworthiness* of information sources.

This paper proposes an argument-based system that allows an agent to reason about her 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.

Categories and Subject Descriptors

I.2.3 [**Deduction and Theorem Proving**]: Nonmonotonic reasoning and belief revision; I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents

General Terms

Human Factors, Theory

Keywords

Dialog, Trust, Argumentation.

1. INTRODUCTION

Since the seminal book by Walton and Krabbe [31] 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. [4]), inquiry (e.g. [6]), negotiation (e.g. [27]) and deliberation (e.g. [23]). Besides, 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. They rather assume that all agents are trustworthy. Indeed, agents 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 (e.g., [9, 10, 17, 21, 25]). This would mean that in existing work, agents may accept incorrect claims and may make deals with unreliable agents.

This paper fills the gap by proposing an argumentation system for reasoning about different kinds of beliefs including beliefs about trust in information sources. The system fulfills thus three tasks. It states whether: i) to believe in a given statement, ii) to trust or not a given source, and iii) to accept or not an information/offer received from a source. The system considers a rich notion of trust as proposed in [12, 13]. Indeed, one may trust in different properties of an agent, namely her *validity*, *completeness*, *sincerity*, *cooperativity*, *competence* and *vigilance*. Besides, 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 beliefs base which is encoded in modal logic and which contains formulas expressing information about the environment (like the window is closed) and information about trust (e.g., agent i trusts in the sincerity of agent j). It builds two types of arguments: arguments in favor of trusting a given source of information and arguments in favour 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 paper is structured as follows: Section 2 introduces the logical formalism that is used for representing beliefs. Section 3 defines six different forms of trust in information sources. Section 4 defines the argumentation system for reasoning about trust information, and Section 5 investigates its properties. Section 6 compares our proposal with existing work on modelling trust. The last section is devoted to some concluding remarks and perspectives.

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

2. LOGICAL FORMALISM

This section introduces the logical formalism (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:

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

$\mathcal{L}$ 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 AT and i and j range over AG. 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 ϕ holds.
- $\mathrm{Inf}_{j,i}\phi$: agent j has informed agent i that ϕ holds.

The axiomatics of the logic is the axiomatics of a Propositional multi Modal Logic (see [11]). In addition to the axiomatics of Classical Propositional Calculus we have the following axiom schemas and inference rules.

(K) $\quad \mathrm{Bel}_i(\phi \to \psi) \to (\mathrm{Bel}_i\phi \to \mathrm{Bel}_i\psi)$

(D) $\quad \neg(\mathrm{Bel}_i\phi \wedge \mathrm{Bel}_i\neg\phi)$

(Nec) $\quad$ If $\vdash \phi$, then $\vdash \mathrm{Bel}_i\phi$

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

The modal operator Inf obeys the following axiom schemas:

(EQV) $\quad$ If $\vdash \phi \leftrightarrow \psi$, then $\vdash \mathrm{Inf}_{j,i}\phi \leftrightarrow \mathrm{Inf}_{j,i}\psi$

(CONJ) $\quad \mathrm{Inf}_{j,i}\phi \wedge \mathrm{Inf}_{j,i}\psi \to \mathrm{Inf}_{j,i}(\phi \wedge \psi)$

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.

In the sequel, the symbol $\vdash$ refers to the consequence operator of the formalism. Besides, a beliefs base is a subset of $\mathcal{L}$ which contains the beliefs of a given agent $i \in$ AG.

3. 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 [12, 13], 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*: the truster (i) believes that if he is informed by the trustee (j) about some proposition, then the trustee believes that this proposition is true meaning that he does not lie. Generally, patients trust in the sincerity of their doctors. Formally:

$$\mathrm{TrustSinc}(i, j, \phi) \overset{\mathrm{def}}{=} \mathrm{Bel}_i(\mathrm{Inf}_{j,i}\phi \to \mathrm{Bel}_j\phi)$$

It is worth mentioning that the fact that an agent i believes in the sincerity of another agent j regarding proposition ϕ does not mean that i believes ϕ. 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*: the truster (i) believes that if he is informed by the trustee (j) about some proposition, then this proposition is true. Generally, a child trusts in validity of his father and thinks that any claim made by the father is necessarily true.

$$\mathrm{TrustVal}(i, j, \phi) \overset{\mathrm{def}}{=} \mathrm{Bel}_i(\mathrm{Inf}_{j,i}\phi \to \phi)$$

Trust in *completeness*: the truster believes that if some proposition is true, then he is informed by the trustee about this proposition. For instance, the inhabitants of a building trust in the completeness of the caretaker of the building. They believe that if the elevator is out of service, they will be informed.

$$\mathrm{TrustCmp}(i, j, \phi) \overset{\mathrm{def}}{=} \mathrm{Bel}_i(\phi \to \mathrm{Inf}_{j,i}\phi)$$

Trust in *cooperativity*: the truster believes that if the trustee believes that some proposition is true, then he is informed by the trustee about this proposition. This is important in information-seeking like dialogs where agents ask questions in order to elicit information from other sources.

$$\mathrm{TrustCoop}(i, j, \phi) \overset{\mathrm{def}}{=} \mathrm{Bel}_i(\mathrm{Bel}_j\phi \to \mathrm{Inf}_{j,i}\phi)$$

Trust in *competence*: the truster believes that if the trustee believes that some proposition is true, then this proposition is true. For instance, a patient trusts in the competence of his doctor and thinks that the diagnosis made by the doctor is necessarily true.

$$\mathrm{TrustComp}(i, j, \phi) \overset{\mathrm{def}}{=} \mathrm{Bel}_i(\mathrm{Bel}_j\phi \to \phi)$$

Trust in *vigilance*: the truster believes that if some proposition is true, then the trustee believes that this proposition is true, i.e., he is aware of the proposition.

$$\mathrm{TrustVigi}(i, j, \phi) \overset{\mathrm{def}}{=} \mathrm{Bel}_i(\phi \to \mathrm{Bel}_j\phi)$$

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 and g_2. This is represented by two formulas: $\mathrm{Bel}_p(\mathrm{Bel}_d g_1 \to g_1)$ and $\mathrm{Bel}_p(\mathrm{Bel}_d g_2 \to g_2)$.

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

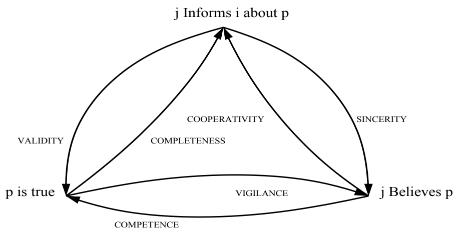


Figure 1: Relationships between believing, informing and truth.

It is also clear that the six formulas are elements of $\mathcal{L}$.

Note that 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 to inform 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:

$$\vdash \text{TrustSinc}(i, j, \phi) \wedge \text{TrustComp}(i, j, \phi) \rightarrow \text{TrustVal}(i, j, \phi)$$

$$\vdash \text{TrustVigi}(i, j, \phi) \wedge \text{TrustCoop}(i, j, \phi) \rightarrow \text{TrustCmp}(i, j, \phi)$$

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

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

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

(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 ϕ 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 ϕ 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.

4. TRUST SUPPORTED BY ARGUMENTS

This section introduces an argumentation system for reasoning about the different kinds of beliefs an agent i may have. Starting from 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 ϕ 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 a 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 ϕ and $\neg\phi$ hold.

The system is a logical instantiation of the abstract framework proposed by Dung in his seminal paper [15]. 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 1. *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 $H' \vdash h$*

H is called the support *of the argument and h its* conclusion. *$\mathtt{Arg}(\mathcal{K}_i)$ 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 1. *Assume the following beliefs base of agent i:*

$$\mathcal{K}_i = \begin{cases} Bel_i(\delta) \\ Bel_i(Inf_{j,i}\phi) \\ Bel_i(\neg Inf_{k,i}\varphi) \\ Bel_i(Inf_{j,i}\phi \rightarrow Bel_j\phi) \\ Bel_i(\varphi \rightarrow Inf_{k,i}\varphi) \end{cases}$$

From this base, an infinite number of arguments are built including the following ones:

1. *$(\{Bel_i(\delta)\}, Bel_i(\delta))$*
2. *$(\{Bel_i(Inf_{j,i}\phi)\}, Bel_i(Inf_{j,i}\phi))$*
3. *$(\{Bel_i(\neg Inf_{k,i}\varphi)\}, Bel_i(\neg Inf_{k,i}\varphi))$*
4. *$(\{Bel_i(Inf_{j,i}\phi \rightarrow Bel_j\phi), Bel_i(Inf_{j,i}\phi)\}, Bel_i(Bel_j\phi))$*
5. *$(\{Bel_i(\varphi \rightarrow Inf_{k,i}\varphi), Bel_i(\neg Inf_{k,i}\varphi)\}, Bel_i\neg\varphi)$*

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 an agent may receive new information from other entities.

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

EXAMPLE 2. *Assume the following base:*

$$\mathcal{K}_i = \begin{cases} Bel_i(\varphi) \to TrustSinc(i,j,\phi) \\ TrustVal(i,k,\varphi) \\ Bel_i(Inf_{k,i}\varphi) \end{cases}$$

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 φ stands for ”j makes fair reviews” and ϕ for ”j makes a fair review for paper ID x”. Examples of arguments that are built from this base are the following ones:

1. $(\{Bel_i(Inf_{k,i}\varphi)\}, Bel_i(Inf_{k,i}\varphi))$
2. $(\{Bel_i(Inf_{k,i}\varphi),\ TrustVal(i,k,\varphi)\}, Bel_i\varphi)$
3. $(\{Bel_i(Inf_{k,i}\varphi),\ TrustVal(i,k,\varphi),$ $Bel_i\varphi \to TrustSinc(i,j,\phi)\},\ TrustSinc(i,j,\phi))$

The argument (3) is in favor of trusting in the sincerity of agent j on proposition ϕ.

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 [19] for a summary of relations proposed for propositional frameworks). Some of them, like the well-known *rebutting*, are symmetric. However, it was shown in [2] that any argumentation framework which is grounded on a Tarskian logic ([29]) and uses a symmetric attack relation may violate the rationality postulates proposed in [8], 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, in what follows we avoid symmetric relations. We discuss next various forms of attacks. The first one is the so-called *assumption-attack* proposed in [16]. It consists of weakening an argument by undermining one of its premises (i.e., an element of its support).

DEFINITION 2. *Let $(H,h),(H',h')$ be two arguments of $\texttt{Arg}(\mathcal{K}_i)$. (H,h) assumption-attacks (H',h') iff there exists $h'' \in H'$ such that $h = Bel_i\phi$ and $h'' = Bel_i\neg\phi$.*

Let us illustrate this relation on the following example.

EXAMPLE 3. *Let us consider the following base:*

$$\mathcal{K}_i = \begin{cases} Bel_i(Inf_{j,i}\phi \to Bel_j\phi) \\ Bel_i(Inf_{j,i}\phi \to \phi) \\ Bel_i(Inf_{j,i}\phi) \\ Bel_i(\neg\phi) \end{cases}$$

The argument $(\{Bel_i(Inf_{j,i}\phi \to \phi), Bel_i(\neg\phi)\}, Bel_i(\neg Inf_{j,i}\phi))$ assumption attacks the argument $(\{Bel_i(Inf_{j,i}\phi \to Bel_j\phi),\ Bel_i(Inf_{j,i}\phi)\},\ Bel_i(Bel_j\phi))$.

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 3. *Let $(H,h),(H',h')$ be two arguments of $\texttt{Arg}(\mathcal{K}_i)$. (H,h) sinc-attacks (H',h') iff $h = Bel_i(Inf_{j,i}\varphi \wedge \neg Bel_j\varphi)$ and $TrustSinc(i,j,\phi) \in H'$.*

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. *Let $(H,h),(H',h')$ be two arguments of $\texttt{Arg}(\mathcal{K}_i)$. (H,h) val-attacks (H',h') iff $h = Bel_i(Inf_{j,i}\varphi \wedge \neg\varphi)$ and $TrustVal(i,j,\phi) \in H'$.*

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 5. *Let $(H,h),(H',h')$ be two arguments of $\texttt{Arg}(\mathcal{K}_i)$. (H,h) com-attacks (H',h') iff $h = Bel_i(\varphi \wedge \neg Inf_{j,i}\varphi)$ and $TrustCmp(i,j,\phi) \in H'$.*

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 6. *Let $(H,h),(H',h')$ be two arguments of $\texttt{Arg}(\mathcal{K}_i)$. (H,h) coop-attacks (H',h') iff $h = Bel_i(Bel_j\varphi \wedge \neg Inf_{j,i}\varphi)$ and $TrustCoop(i,j,\phi) \in H'$.*

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 7. *Let $(H,h),(H',h')$ be two arguments of $\texttt{Arg}(\mathcal{K}_i)$. (H,h) comp-attacks (H',h') iff $h = Bel_i(Bel_j\varphi \wedge \neg\varphi)$ and $TrustComp(i,j,\phi) \in H'$.*

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

DEFINITION 8. *Let $(H,h),(H',h')$ be two arguments of $\texttt{Arg}(\mathcal{K}_i)$. (H,h) vigi-attacks (H',h') iff $h = Bel_i(\varphi \wedge \neg Bel_j\varphi)$ and $TrustVigi(i,j,\phi) \in H'$.*

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. *Let us consider the following base:*

$$\mathcal{K}_i = \begin{cases} Bel_i(Inf_{j,i}\phi \to Bel_j\phi) \\ Bel_i(Inf_{j,i}\varphi) \\ Bel_i(\neg Bel_j\varphi) \end{cases}$$

Assume that ϕ stands for 'The weather is cloudy' and φ stands for 'People pay few taxes'. Note that the base $\mathcal{K}_i$ is consistent. However, the argument $(\{Bel_i(Inf_{j,i}\varphi), Bel_i(\neg Bel_j\varphi)\}, Bel_i(Inf_{j,i}\varphi \wedge \neg Bel_j\varphi))$ sinc-attacks the argument $(\{Bel_i(Inf_{j,i}\phi \to Bel_j\phi)\}, Bel_i(Inf_{j,i}\phi \to Bel_j\phi))$.

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

DEFINITION 9. *Let (H,h) and (H',h') be two arguments of $\texttt{Arg}(\mathcal{K}_i)$. (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 (Cont) It is easy to check that there is only one attack between arguments of $\mathtt{Arg}(\mathcal{K}_i)$: $(\{\mathrm{Bel}_i(\mathrm{Inf}_{j,i}\varphi), \mathrm{Bel}_i(\neg\mathrm{Bel}_j\varphi)\}, \mathrm{Bel}_i(\mathrm{Inf}_{j,i}\varphi \wedge \neg\mathrm{Bel}_j\varphi)) \ \Re \ (\{\mathrm{Bel}_i(\mathrm{Inf}_{j,i}\phi \rightarrow \mathrm{Bel}_j\phi)\}, \mathrm{Bel}_i(\mathrm{Inf}_{j,i}\phi \rightarrow \mathrm{Bel}_j\phi))$. Thus, $\Re$ is not symmetric.

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

Example 5. *Let us consider the following base:*

$$\mathcal{K}_i = \left\{ \begin{array}{l} \mathit{TrustSinc}(i, j, \phi) \\ \mathit{Bel}_i((\mathit{Inf}_{j,i}\phi \rightarrow \mathit{Bel}_j\phi) \rightarrow \mathit{Bel}_i(\neg \mathit{Bel}_j\varphi)) \\ \mathit{Bel}_i(\mathit{Inf}_{j,i}\varphi) \end{array} \right.$$

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

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

Definition 10. *An argumentation system built over a beliefs base* $\mathcal{K}_i$ *is a pair* $\mathcal{T} = (\mathtt{Arg}(\mathcal{K}_i), \Re)$ *where* $\Re \subseteq \mathtt{Arg}(\mathcal{K}_i) \times \mathtt{Arg}(\mathcal{K}_i)$ *is as given in Definition 9.*

Since arguments may be conflicting, it is important to define the acceptable ones. For that purpose, we use the stable semantics proposed by Dung in [15]. This semantics allows to partition the powerset of the set of arguments into two sets: stable extensions and non-extensions. An extension is a set of arguments that are acceptable together. It represents thus a coherent point of view.

Definition 11. *Let* $\mathcal{T} = (\mathtt{Arg}(\mathcal{K}_i), \Re)$ *be an argumentation system built over a beliefs base* $\mathcal{K}_i$ *and* $\mathcal{E} \subseteq \mathtt{Arg}(\mathcal{K}_i)$*.* $\mathcal{E}$ *is a* stable *extension iff:*

- $\nexists a, b \in \mathcal{E}$ *such that* $(a, b) \in \Re$*.*
- $\mathcal{E}$ *attacks*[2] *any argument in* $\mathtt{Arg}(\mathcal{K}_i) \setminus \mathcal{E}$*.*

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

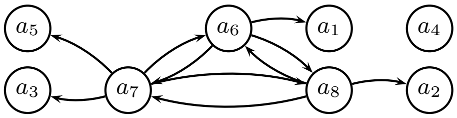
The extensions are used in order to define the inferences to be drawn from the beliefs base $\mathcal{K}_i$ of agent i. These inferences represent what agent i *should believe* according to the available information. The idea is that a formula is inferred if it is supported by at least one argument in every extension. Note that the argument *needs not to be the same in all the extensions.*

Definition 12. *Let* $\mathcal{T} = (\mathtt{Arg}(\mathcal{K}_i), \Re)$ *be an argumentation system built over a beliefs base* $\mathcal{K}_i$ *and* $\mathtt{Ext}(\mathcal{T})$ *its set of stable extensions. A formula* $\phi \in \mathcal{L}$ *is* inferred *from* $\mathcal{K}_i$ *iff for all* $\mathcal{E} \in \mathtt{Ext}(\mathcal{T})$*, there exists* $(H, \phi) \in \mathcal{E}$*.*
$\mathtt{Output}(\mathcal{T})$ *denotes the set of all beliefs inferred from* $\mathcal{K}_i$ *using system* $\mathcal{T}$*.*

Example 3 (Cont) Let us consider the beliefs base $\mathcal{K}_i$ of agent i. The set $\mathtt{Arg}(\mathcal{K}_i)$ of arguments is infinite. It contains among others the following arguments:

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

The following figure summarizes the attacks between the eight arguments:



It can be checked that the argumentation system $\mathcal{T} = (\mathtt{Arg}(\mathcal{K}_i), \Re)$ has three stable extensions. Note that we do not provide the complete result since $\mathtt{Arg}(\mathcal{K}_i)$ 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, $a_1 \notin \mathcal{E}_1$.

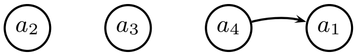
- $\mathcal{E}_1 = \{a_2, a_3, a_4, a_5, a_6, \ldots\}$
- $\mathcal{E}_2 = \{a_1, a_2, a_4, a_7, \ldots\}$
- $\mathcal{E}_3 = \{a_1, a_3, a_4, a_5, a_8, \ldots\}$.

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

Example 4 (Cont)
The table below shows some arguments that may be built from $\mathcal{K}_i$.

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

The following figure summarizes the attacks between the four arguments:



[2] A set $\mathcal{E}$ of arguments attacks an argument a iff there exists $b \in \mathcal{E}$ such that $(b, a) \in \Re$.

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

5. PROPERTIES OF THE SYSTEM

In this section, we investigate the properties of the proposed model.

Remember that a beliefs 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 [8] for rule-based logics and generalized later in [1] for Tarskian logics.

PROPOSITION 1. *Let $\mathcal{T} = (\texttt{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$ and $\texttt{Ext}(\mathcal{T})$ its set of stable extensions. For all $\mathcal{E} \in \texttt{Ext}(\mathcal{T})$, 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 $\mathcal{T} = (\texttt{Arg}(\mathcal{K}_i), \Re)$. 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 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 $\texttt{Output}(\mathcal{T})$ is also consistent.

PROPOSITION 2. *Let $\mathcal{T} = (\texttt{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$. The set $\texttt{Output}(\mathcal{T})$ is consistent.*

PROOF. From Definition 12, it follows that $\texttt{Output}(\mathcal{T}) \subseteq \{h \mid \exists(H,h) \in \mathcal{E}\}$ for any $\mathcal{E} \in \texttt{Ext}(\mathcal{T})$. Since $\{h \mid \exists(H,h) \in \mathcal{E}\}$ is consistent then so is $\texttt{Output}(\mathcal{T})$. □

The next property concerns another rationality postulate in [1] 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 3. *Let $\mathcal{T} = (\texttt{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$. For all $\mathcal{E} \in \texttt{Ext}(\mathcal{T})$, if $(H,h) \in \mathcal{E}$ then for all $(H',h') \in \texttt{Arg}(\mathcal{K}_i)$ such that $H' \subseteq H$, it holds that $(H',h') \in \mathcal{E}$.*

PROOF. Let $\mathcal{E}$ be a stable extension of $\mathcal{T} = (\texttt{Arg}(\mathcal{K}_i), \Re)$. Let $(H,h) \in \mathcal{E}$ and $(H',h') \in \texttt{Arg}(\mathcal{K}_i)$ 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 [1] 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 $X \subseteq \mathcal{L}$, $\texttt{CN}(X) = \{\phi \in \mathcal{L} \mid X \vdash \phi\}$.

PROPOSITION 4. *Let $\mathcal{T} = (\texttt{Arg}(\mathcal{K}_i), \Re)$ be an argumentation system built over a beliefs base $\mathcal{K}_i$ and $\texttt{Ext}(\mathcal{T})$ its set of stable extensions. For all $\mathcal{E} \in \texttt{Ext}(\mathcal{T})$, $\{h \mid \exists(H,h) \in \mathcal{E}\} = \texttt{CN}(\{h \mid \exists(H,h) \in \mathcal{E}\})$.*

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

1) $h \in \texttt{CN}(\emptyset)$, $(\emptyset, h) \in \texttt{Arg}(\mathcal{K}_i)$ 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) $h \notin \texttt{CN}(\emptyset)$ and $\exists S \subseteq \bigcup_{(H_k,h_k)\in\mathcal{E}} H_k$ such that $(S,h) \in \texttt{Arg}(\mathcal{K}_i)$ since $\bigcup_{(H_k,h_k)\in\mathcal{E}} H_k$ is consistent (see Proposition 1). 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 $\texttt{Output}(\mathcal{T})$ is closed under $\vdash$.

Proposition 5. *Let* $\mathcal{T} = (\mathtt{Arg}(\mathcal{K}_i), \Re)$ *be an argumentation system built over a beliefs base* $\mathcal{K}_i$ *such that* $\mathtt{Ext}(\mathcal{T}) \neq \emptyset$*. It holds that* $\mathtt{Output}(\mathcal{T}) = \mathtt{CN}(\mathtt{Output}(\mathcal{T}))$*.*

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

Assume now that $h \in \mathtt{CN}(\mathtt{Output}(\mathcal{T}))$ and $h \notin \mathtt{Output}(\mathcal{T})$. Then, $\exists h_1, \ldots, h_n \in \mathtt{Output}(\mathcal{T})$ such that $h \in \mathtt{CN}(\{h_1, \ldots, h_n\})$. Besides, h_1, …, $h_n \in \bigcap_{\mathcal{E}_k \in \mathtt{Ext}(\mathcal{T})} \{\phi \mid \exists(H, \phi) \in \mathcal{E}_k\}$. From monotonicity of $\mathtt{CN}$, it follows that: $\mathtt{CN}(\{h_1, \ldots, h_n\}) \subseteq \mathtt{CN}(\bigcap_{\mathcal{E}_k \in \mathtt{Ext}(\mathcal{T})} \{\phi \mid \exists(H, \phi) \in \mathcal{E}_k\})$. It holds also that $h \in \mathtt{CN}(\{\phi \mid \exists(H, \phi) \in \mathcal{E}_1\}) \cap \ldots \cap \mathtt{CN}(\{\phi \mid \exists(H, \phi) \in \mathcal{E}_n\})$. From Proposition 4, $h \in \{\phi \mid \exists(H, \phi) \in \mathcal{E}_1\} \cap \ldots \cap \{\phi \mid \exists(H, \phi) \in \mathcal{E}_n\}$. Consequently, $h \in \mathtt{Output}(\mathcal{T})$. □

This means, for instance, that if $\text{TrustSinc}(i, j, \phi) \in \mathtt{Output}(\mathcal{T})$ and $\text{Bel}_i(\text{Inf}_{j,i}\phi) \in \mathtt{Output}(\mathcal{T})$, then $\text{Bel}_i(\text{Bel}_j\phi) \in \mathtt{Output}(\mathcal{T})$.

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. For instance, in [18] trust is defined as a subjective probability by which an agent i expects that another agent j performs a given action on which its welfare depends. In [20], 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 ϕ and i trusts the judgment of j on ϕ, then he will also believe ϕ. 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 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 [20], an instantiation of the meta-argumentation model [7] for reasoning about trust in validity was proposed in [30]. 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 [28]. It is as an instantiation of the preference-based argumentation framework proposed in [3] 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 trust 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. Weights are associated with edges expressing 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 [28]. Moreover, our formalism not only uses trust in order to infer new beliefs but also reasons about trust itself and infers beliefs about trust.

More recently, in [24] 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 [26]. 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 [5]. Our formalism does not use statistical data. Moreover, it is an inference model and not a decision making one.

Finally, in [22] 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 [14]. The logical language of this logic contains a predicate $A(t, \phi)$ whose intuitive meaning is that formula ϕ 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.

8. REFERENCES

[1] L. Amgoud. Postulates for logic-based argumentation systems. In *ECAI Workshop on Weighted Logics for AI (WL4AI'12)*, 2012.

[2] L. Amgoud and P. Besnard. Bridging the gap between abstract argumentation systems and logic. In *International Conference on Scalable Uncertainty Management, SUM'09*, pages 12–27, 2009.

[3] L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Art. Intel.*, 34:197–216, 2002.

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

[5] L. Amgoud and H. Prade. Using arguments for making and explaining decisions. *Artificial Intelligence Journal*, 173:413–436, 2009.

[6] E. Black and A. Hunter. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems*, 19(2):173–209, 2009.

[7] G. Boella, D. Gabbay, L. van der Torre, and S. Villata. Meta-argumentation modelling i: Methodology and techniques. *Studia Logica*, 93(2-3):297–355, 2009.

[8] M. Caminada and L. Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence Journal*, 171 (5-6):286–310, 2007.

[9] C. Castelfranchi. Trust: nature and dynamics. In *ACM SIGCHI Italian Chapter International Conference on Computer-Human Interaction*, pages 13–14, 2011.

[10] C. Castelfranchi and R. Falcone. Trust is much more than subjective probability: Mental components and sources of trust. In *HICSS*, 2000.

[11] B. Chellas. *Modal logic: an introduction*. Cambridge University Press, Cambridge, 1980.

[12] R. Demolombe. To trust information sources: a proposal for a modal logical framework. In C. Castelfranchi and Y-H. Tan, editor, *Proc. of the Workshop on Deception, Fraud and Trust in Agent Societies*, 1998.

[13] R. Demolombe. Reasoning about trust: A formal logical framework. In *Second International Conference on Trust Management, iTrust'04*, pages 291–303, 2004.

[14] R. Demolombe and A. Jones. Reasoning about Topics: towards a formal theory. In *American Association for Artificial Intelligence Fall Symposium*, 1995.

[15] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence Journal*, 77:321–357, 1995.

[16] M. Elvang-Gøransson, J. Fox, and P. Krause. Acceptability of arguments as 'logical uncertainty. In *Proceedings of the 2nd European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'93*, pages 85–90, 1993.

[17] R. Falcone, M. Piunti, M. Venanzi, and C. Castelfranchi. From manifesta to krypta: The relevance of categories for trusting others. *ACM TIST*, 4(2):27, 2013.

[18] D. Gambetta. Can we trust them? *Trust: Making and breaking cooperative relations*, pages 213–238, 1990.

[19] N. Gorogiannis and A. Hunter. Instantiating abstract argumentation with classical logic arguments: Postulates and properties. *Artificial Intelligence Journal*, 175 (9–10):1479–1497, 2011.

[20] C. Liau. Belief, information acquisition, and trust in multi-agent systems–a modal logic formulation. *Artificial Intelligence Journal*, 149(1):31–60, 2003.

[21] S. Marsh. Formalising trust as a computational concept. Technical report, Ph.D. Thesis, University of Stirling, 1994.

[22] P. Matt, M. Morge, and F. Toni. Combining statistics and arguments to compute trust. In *International Conference on Autonomous Agents and Multiagent Systems, AAMAS'2010*, pages 209–216, 2010.

[23] P. McBurney, D. Hitchcock, and S. Parsons. The eightfold way of deliberation dialogue. *International Journal of Intelligent Systems*, 22(1):95–132, 2007.

[24] S. Parsons, K. Atkinson, K. Haigh, K. Levitt, P. McBurney, J. Rowe, M. Singh, and E. Sklar. Argument schemes for reasoning about trust. In *Computational Models of Argument, COMMA'12*, pages 430–441, 2012.

[25] J. Shi, G. Bochmann, and C. Adams. A trust model with statistical foundation. *IFIP Advances in Information and Communication Technology*, 173:145–158, 2005.

[26] R. Stranders, M. de Weerdt, and C. Witteveen. Fuzzy argumentation for trust. In *International Workshop on Computational Logic in Multi-Agent Systems, CLIMA'07*, pages 214–230, 2007.

[27] K. Sycara. Persuasive argumentation in negotiation. *Theory and Decision*, 28:203–242, 1990.

[28] Y. Tang, K. Cai, P. McBurney, E. Sklar, and S. Parsons. Using argumentation to reason about trust and belief. *Journal of Logic and Computation*, 22(5):979–1018, 2012.

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

[30] S. Villata, G. Boella, D. Gabbay, and L. van der Torre. Arguing about the trustworthiness of the information sources. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU'11*, pages 74–85, 2011.

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