A socio-cognitive model of trust using argumentation theory

A Socio-Cognitive Model of Trust using Argumentation Theory

Serena Villata[a], Guido Boella[b], Dov M. Gabbay[c], Leendert van der Torre[d]

[a]*INRIA Sophia Antipolis, France*
[b]*University of Torino, Italia*
[c]*King's College, United Kingdom*
[d]*University of Luxembourg, Luxembourg*

Abstract

Trust is used to minimize the uncertainty in the interactions of agents especially in case of conflicting information from different sources. Besides conflicts among information there can be conflicts about the trust attributed to the information sources. In this paper, we discuss how to reason about trust by using argumentation theory, so to express also the possibly conflicting motivations about trust and distrust. The methodology of meta-argumentation allows us to model both information and information sources as arguments and to argue about them. First, we present a model for representing evidence provided in support of the sources' arguments to represent the need of a trusted source to believe the information, and we show how to model the information sources in a way that it can be argued if they should considered untrustworthy or not. Second, we provide a focused representation of trust about the sources in which trust concerns not only the sources but also the information items and the relation with other information. Third, we introduce the feedback on the trustworthiness of the sources and the information items they propose, such that an attack to the trustworthiness of the items feeds back on the source's one. Finally, we distinguish two dimensions of trust, namely competence and sincerity, and we present a formal way to express those dimensions, only informally described in the socio-cognitive models of trust.

Keywords: Trust, meta-argumentation

1. Introduction

Trust is a mechanism for managing uncertain information in decision making, taking into account also the sources besides the content of information only. In

their interactions, agents have to reason whether they should trust or not the other sources of information, and on the extent to which they trust those other sources. This is important, for example, in medical contexts, where doctors have to inform the patient of the pro and con evidence from different sources concerning some treatment, in decision support systems where the user is not satisfied by an answer without explanations, or in trials where judges have to specify the motivations about which conflicting evidence they trust.

A cognitive analysis of trust is fundamental for distinguishing between internal and external attribution which predict very different strategies for building or increasing trust, for founding mechanisms of reputation, persuasion, and argumentation in trust building [1].

In this paper, we start from the cognitive model of trust introduced by Castelfranchi and Falcone [1], and we present a cognitive model of conflicts in trust using argumentation. In particular, the reasoning process addressed by the agents concerning the extent to which they trust the other information sources leads to the emergence not only of conflicts among the information but also of the conflicts among the sources. Since argumentation is a mechanism to reason about conflicting information [2] it seems the suitable methodology for reason about trust. When two pieces of information coming from different sources are conflicting, they can be seen as two arguments attacking each other. When an information source explicitly expresses a negative evaluation of the trustworthiness of another source, it can be seen as an "attack" to the trustworthiness of the second source modelled as an argument as well. To deal with the dimension of conflict in handling trust, we propose to use argumentation theory, modelling both information and information sources as arguments and arguing about them. In argumentation theory [3], the arguments are considered to be accepted or not depending on the attacks against them. In standard argumentation frameworks, neither the information sources proposing the arguments nor their trustworthiness are considered. In recent years, the area has seen a number of proposals [4, 5, 6, 7, 8, 9] to introduce the trust component in the evaluation process of the arguments. The common drawback of these approaches is that they do not return the intrinsic complexity of the trust notion, as highlighted instead by socio-cognitive models like [1].

The challenge of this work is to use argumentation theory not only to model whether an information source is trusted or not, but also to understand the reasons, modeled under the form of arguments, for trusting the sources in case of conflicts concerning their trustability. This means that we need to distinguish the conflicts about the content of the arguments which are usually specified through an attack relation, and the conflicts about the different opinions of the sources on

the trustworthiness of the other sources. These are two separate reasoning levels, and the challenge is to model both of them using argumentation theory. In particular, we present a way to deal with the conflicts about trust using Dung's abstract argumentation framework [10]. It is not obvious how to model in a Dung argumentation framework the trust about arguments and the conflicts about sources. A Dung argumentation framework can be instantiated by the arguments and attacks defined by a knowledge base. The knowledge base inferences are defined in terms of the claims of the justified arguments, e.g., the ASPIC+ framework [11] instantiates Dung frameworks with accounts of the structure of arguments, the nature of attack and the use of preferences. In such a kind of framework, arguments are instantiated by sentences of a single knowledge base, without reference to the information sources. The only possibility is to include sources and trust inside the content of the argument. This makes it difficult to distinguish between the object level concerning content of information and the meta-level concerning trust, sources and the conflicts among them. In reasoning about trust, the information about the trustworthiness relations among the sources are meta-level information, and they cannot be inserted directly into the arguments of the framework. They influence the behavior of the framework in the sense that they lead to further conflicts among the sources and their information items, i.e., what the sources claim.

The following example presents informally the opinions of several witnesses during a trial, illustrating conflicts about trust among the sources and not only among the pieces of information they provide, where the external evaluator is the judge:

- *Witness1: I suspect that the man killed his boss in Rome. (a)*
- *Witness1: But his car was broken, thus he could not reach the crime scene. (b)*
- *Witness2: Witness1 is a compulsive liar. (c)*
- *Witness3: I repaired the suspect's car at 12pm of the crime day. (d)*
- *Witness4: I believe that Witness3 is not able to repair that kind of car. (e)*
- *Witness5: The suspect has another car. (f)*
- *Witness6: Witness5 saw that the suspect parked 2 cars in my underground parking garage 3 weeks ago. (g)*
- *Witness2: Witness5 was on holidays 3 weeks ago. (h)*
- *Witness7: Witness5 cannot go on holidays because of his working contract. (i)*

- *Witness3: Witness7 is not competent about the working contracts of the underground parking garage. (l)*
- *Witness1: Witness7 does not really think that Witness5 cannot go on holidays because of his working contract. (m)*

In these sentences, different kinds of conflicts are highlighted among the sources concerning their trustability. What we call the object level is illustrated by the arguments (a) and (b): Witness1 would believe the suspect is the murderer but he explains that another argument (the car was broken) prevents this conclusion. Thus argument (b) attacks argument (a) since they are conflicting. But attacks can concern also the trustability of sources, once this aspect is modelled in terms of arguments (meta-arguments) as well. First, the sources can attack the trustworthiness of the other sources, see, e.g., argument (c) attacking the trustworthiness of Wintess1. Second, we must model the connection between the argument about the trustability of Witness1 and the arguments (a) and (b) - as well as the attack between the two arguments - he advances. The sources must be modelled as evidence in support of their arguments or attacks, which otherwise should be considered as not acceptable. Moreover, sources can provide evidence also concerning the other sources' arguments, e.g., argument (g) provides evidence for argument (f). Third, while attacks like the one done by argument (c) are addressed against the sources' trustworthiness as a whole represented as an argument, conflicts about trust can be restricted to a particular argument or attack proposed by a source who is not considered untrustworthy in general. E.g., argument (h) expresses concerns about the trustworthiness of argument (g) and not about the source itself. Fourth, conflicts about the trustworthiness of the sources can be further detailed in order to deal with the competence of the sources, e.g., argument (l), and their sincerity, e.g., argument (m). Last, implicit in the example, there is the issue of a feedback between the trustworthiness of the information items proposed by the sources and the sources' trustworthiness when what they said is attacked.

The problem in standard argumentation frameworks [3] is that it is difficult to formalize the example above with sentences from a single knowledge base only, e.g., to model it in ASPIC+ style instantiated argumentation. Moreover, meta-level information such as the distinction about conflicts based on sincerity and those based on competence cannot be represented in those frameworks. These two trust dimensions might be independently evaluated in the argumentation process: Bob's sincerity/honesty (Alice believes that Bob has told her the truth) vs. Bob's competence (Alice trusts the judgment of Bob if he is expert). Finally, it has to be modeled the fact that attacking Bob's argument means attacking Bob and his

credibility and trustworthiness as source. This is fundamental, both in the case in which it is intentional and it is the real objective of the attack, or when it is not intended but is a consequence of the invalidation of the arguments. This is because of the bidirectional link between the source and its information items: the provided item is more or less believable on the basis of the source trustworthiness, but the invalidation of the item feedbacks on the source's credibility.

In this paper, we address the following research question:

- How to model the socio-cognitive aspects of trust using argumentation theory?

The research question breaks down into the following subquestions:

1. How to represent the information sources and attack their trustworthiness?
2. How to represent pro and con evidence, as done in Carneades [12]?
3. How to attack the sources' trustworthiness about single information items?
4. How to represent the trust feedback between the sources and their information items?
5. How to distinguish the two dimensions of trust, i.e., sincerity and competence?

To answer the research questions, we propose meta-argumentation [13, 14, 15, 16, 17]. Meta-argumentation provides a way to instantiate abstract arguments, i.e., abstract arguments are treated as meta-arguments: arguments about other arguments. It allows us not only to reason about arguments such as sentences from a knowledge base indexed by the information source, but also to introduce in the framework, at the meta-level, other instances like arguments about the trustworthiness of sources. The advantage of adopting meta-argumentation is that we do not extend Dung's framework in order to introduce trust but we instantiate his theory with meta-arguments. For a further discussion about meta-argumentation, see Villata [17].

The sources are introduced into the argumentation framework under the form of meta-arguments of the kind "*agent i is trustable*". An attack to the trustworthiness of a source is modeled as an attack to the meta-argument "*agent i is trustable*". Similarly, in meta-argumentation, both arguments and attacks are represented as meta-arguments, thus allowing arguments to attack attacks.

Each source supports the information items it proposes via meta-arguments which represent the need of evidence to make an argument acceptable. Each argument simply "put on the table" is considered unacceptable if no sources provide an evidence supporting it by being considered trustable.

The information sources propose information items, i.e., arguments, and attacks among these arguments. An attack to the trustworthiness of an item or an attack is modeled as an attack in the meta-level to the evidence provided by the source for that item.

The feedback from the sources to the information items and back is modeled again by introducing new meta-arguments, and the attacks among them. These meta-arguments models a sort of threshold such that if a number of attackers of the information items proposed by a source are accepted, i.e., trustable, thus the attacked source cannot be considered trustworthy.

Finally, the two dimensions of sincerity and competence are modeled using a meta-argument of the kind "i is believed by source s" representing the fact that argument *i* is believed by the source, and thus the source is sincere in proposing *i*. This meta-argument supports the "real" meta-argument which models argument *i* in the meta-level. An attack towards the source's sincerity is modeled as an attack towards the meta-argument representing the believed argument while an attack to the competence is directed towards the support the "believed" meta-argument provides to the "content" meta-argument.

Note that we do not claim that argumentation is the only way to model trust, but we underline that, when the sources argue, they are strongly influenced by the trustworthiness they assign to the other sources. Moreover, we do not assign a numerical value associated to trust, because we are more interested in reasoning about the motivations of the sources, e.g., in the case of Witness1 we have that he explains that he does not believe *a* and that this is due to argument *b*. Finally, we do not treat converging and diverging beliefs sources, and the source's subjective uncertainty [1]. This is left as future work.

The paper follows the research questions. After a brief introduction on meta-argumentation, we describe our cognitive model of trust.

2. Argumentation theory

2.1. Abstract argumentation

A Dung-style framework is based on a binary *attack* relation among arguments, which are abstract entities whose role is determined only by their relation with the other arguments. A Dung-style argumentation framework [10] aims at representing conflicts among elements called *arguments*. It allows to reason about these conflicts in order to detect, starting by a set of arguments and the conflicts among them, which are those arguments which can be considered acceptable. The

acceptable arguments are arguments which are considered as believable by an external evaluator, who has a full knowledge of the argumentation framework.

Definition 1 (Argumentation framework)**.** *An argumentation framework (AF) is a tuple $\langle A, \rightarrow \rangle$ where A is a finite set of elements called arguments and $\rightarrow$ is a binary relation called attack defined on $A \times A$.*

Definition 2 (Defence)**.** *Let $\langle A, \rightarrow \rangle$ be an argumentation framework. Let $\mathscr{S} \subseteq A$. $\mathscr{S}$ defends a if $\forall b \in A$ such that $b \rightarrow a$, $\exists c \in \mathscr{S}$ such that $c \rightarrow b$.*

All Dung's semantics are based on the notion of defence. A semantics of an argumentation theory consists of a conflict free set of arguments, i.e., a set of arguments that does not contain an argument attacking another argument in the set.

Definition 3 (Conflict-free)**.** *Let $\langle A, \rightarrow \rangle$ be an argumentation framework. The set $\mathscr{S} \subseteq A$ is conflict-free if and only if there are no $a, b \in \mathscr{S}$ such that $a \rightarrow b$.*

Like Baroni and Giacomin [18], we use a function $\mathscr{E}$ mapping an argumentation framework $\langle A, \rightarrow \rangle$ to its set of extensions, i.e., to a set of sets of arguments. Since they do not give a name to the function $\mathscr{E}$, and it maps argumentation frameworks to the set of accepted arguments, we call $\mathscr{E}$ the *acceptance function*.

Definition 4. *Let $\mathscr{U}$ be the universe of arguments. An acceptance function $\mathscr{E} : 2^{\mathscr{U}} \times 2^{\mathscr{U} \times \mathscr{U}} \rightarrow 2^{2^{\mathscr{U}}}$ is a partial function which is defined for each argumentation framework $\langle A, \rightarrow \rangle$ with finite $A \subseteq \mathscr{U}$ and $\rightarrow \subseteq A \times A$, and maps an argumentation framework $\langle A, \rightarrow \rangle$ to sets of subsets of A: $\mathscr{E}(\langle A, \rightarrow \rangle) \subseteq 2^A$.*

The following definition summarizes the most widely used acceptability semantics of arguments [10]. Which semantics is most appropriate in which circumstances depends on the application domain of the argumentation theory.

Definition 5 (Acceptability semantics)**.** *Let $AF = \langle A, \rightarrow \rangle$ be an argumentation framework. Let $\mathscr{S} \subseteq A$.*

- *$\mathscr{S}$ is an* admissible *extension if and only if it is conflict-free and defends all its elements.*
- *$\mathscr{S}$ is a* complete extension *if and only if it is conflict-free and we have $\mathscr{S} = \{a \mid \mathscr{S} \, defends \, a\}$.*

- $\mathscr{S}$ *is a* grounded extension *of AF if and only if* $\mathscr{S}$ *is the smallest (for set inclusion) complete extension of AF.*
- $\mathscr{S}$ *is a* preferred extension *of AF if and only if* $\mathscr{S}$ *is maximal (for set inclusion) among admissible extensions of AF.*
- $\mathscr{S}$ *is a* stable extension *of AF if and only if* $\mathscr{S}$ *is conflict-free and attacks all arguments of* $A \setminus \mathscr{S}$.

2.2. Meta-argumentation

Meta-argumentation instantiates Dung's theory with meta-arguments, such that *Dung's theory is used to reason about itself* [19, 15, 17]. Meta-argumentation is a particular way to define mappings from argumentation frameworks to extended argumentation frameworks: arguments are interpreted as meta-arguments, of which some are mapped to "argument a is accepted", $acc(a)$, where a is an abstract argument from the extended argumentation framework EAF. Moreover, auxiliary arguments are introduced to represent, for example, attacks, so that, by being arguments themselves, they can be attacked or attack other arguments. The meta-argumentation methodology is summarized in Figure 2.2.

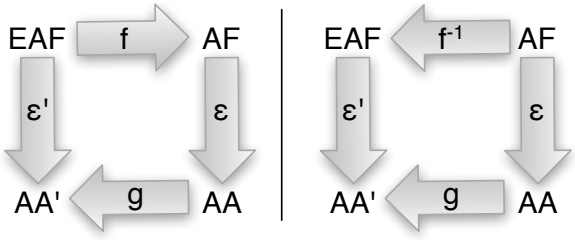

Figure 1: The meta-argumentation methodology workflow.

The function f assigns to each argument a in the EAF, a meta-argument "argument a is accepted" in the basic argumentation framework. The function f^{-1} instantiates an AF with an EAF. We use Dung's acceptance functions $\mathscr{E}$ to find functions $\mathscr{E}'$ between EAFs and the acceptable arguments AA' they return. The acceptable arguments of the meta-argumentation framework are a function of the extended argumentation framework: $AA' = \mathscr{E}'(EAF)$. The transformation function consists of two parts: the function f^{-1}, transforming an argumentation framework to an extended argumentation framework, and a function g which

transforms the acceptable arguments of the argumentation framework into acceptable arguments of the extended argumentation framework. Summarizing, $\mathscr{E}' = \{(f^{-1}(a), g(b)) \mid (a,b) \in \mathscr{E}\}$ and $AA' = \mathscr{E}'(EAF) = g(AA) = g(\mathscr{E}(AF)) = g(\mathscr{E}(f(EAF)))$.

The first step of the meta-argumentation approach is to define the set of extended argumentation frameworks. The second step consists of defining flattening algorithms as a function from this set of EAFs to the set of all basic AF: $f : EAF \rightarrow AF$. The inverse of the flattening is the instantiation of the argumentation framework. See [15, 17] for further details. We define an EAF as a set of partial argumentation frameworks of the sources $\langle A, \langle A_1, \rightarrow_1 \rangle, \ldots, \langle A_n, \rightarrow_n \rangle, \rightarrow \rangle$ [20].

Definition 6. *An extended argumentation framework (EAF) is a tuple* $\langle A, \langle A_1, \rightarrow_1 \rangle, \ldots, \langle A_n, \rightarrow_n \rangle, \rightarrow \rangle$ *where for each source* $1 \leq i \leq n$, $A_i \subseteq A \subseteq \mathscr{U}$ *is a set of arguments,* $\rightarrow$ *is a binary attack relation on* $A \times A$, *and* $\rightarrow_i$ *is a binary relation on* $A_i \times A_i$. *The universe of meta-arguments is* $MU = \{acc(a) \mid a \in \mathscr{U}\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in \mathscr{U}\}$, *where* $X_{a,b}, Y_{a,b}$ *are the meta-arguments corresponding to the attack* $a \rightarrow b$. *The flattening function* f *is given by* $f(EAF) = \langle MA, \longmapsto \rangle$, *where MA is the set of meta-arguments and* $\longmapsto$ *is the meta-attack relation. For a set of arguments* $B \subseteq MU$, *the unflattening function g is given by* $g(B) = \{a \mid acc(a) \in B\}$, *and for sets of subsets of arguments* $AA \subseteq 2^{MU}$, *it is given by* $g(AA) = \{g(B) \mid B \in AA\}$.

Given an acceptance function $\mathscr{E}$ *for an AF, the extensions of accepted arguments of an EAF are given by* $\mathscr{E}'(EAF) = g(\mathscr{E}(f(EAF)))$. *The derived acceptance function* $\mathscr{E}'$ *of the EAF is thus* $\mathscr{E}' = \{(f^{-1}(a), g(b)) \mid (a,b) \in \mathscr{E}\}$. *We say that the source i provides evidence in support of argument a when* $a \in A_i$, *and that the source i supports the attack* $a \rightarrow b$ *when* $a \rightarrow b \in \rightarrow_i$.

Note that the union of all the A_i does not produce A because A contains also those arguments which are not supported by the sources, and are just "put on the table". Definition 7 presents the instantiation of a basic argumentation framework as a set of partial argumentation frameworks of the sources using meta-argumentation.

Definition 7. *Given an* $EAF = \langle A, \langle A_1, \rightarrow_1 \rangle, \ldots, \langle A_n, \rightarrow_n \rangle, \rightarrow \rangle$ *where for each source* $1 \leq i \leq n$, $A_i \subseteq A \subseteq \mathscr{U}$ *is a set of arguments,* $\rightarrow \subseteq A \times A$, *and* $\rightarrow_i \subseteq A_i \times A_i$ *is a binary relation over* A_i. $MA \subseteq MU$ *is* $\{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\}$, *and* $\longmapsto \subseteq MA \times MA$ *is a binary relation on MA such that:* $acc(a) \longmapsto X_{a,b}, X_{a,b} \longmapsto Y_{a,b}, Y_{a,b} \longmapsto acc(b)$ *if and only if there is a source* $1 \leq i \leq n$ *such that* $a,b \in A_i$ *and* $a \rightarrow b \in \rightarrow_i$.

Intuitively, the $X_{a,b}$ auxiliary argument means that the attack $a \rightarrow b$ is "inactive", and the $Y_{a,b}$ auxiliary argument means that the attack is "active". An argument of an EAF is acceptable if and only if it is acceptable in the flattened argumentation framework.

3. Related work

Dix et al. [2] present trust as a major issue concerning the research challenges for argumentation. The question *Which agents are trustworthy?* is important for taking decisions and weighing the arguments of the other agents.

Also Parsons et al. [21] present the provenance of trust as one of the mechanisms to be investigated in argumentation. They claim that a problem, particularly of abstract approaches such as Dung [22], is that they cannot express the provenance of trust, and the fact that argument b is attacked because b is proposed by source s, who is not trustworthy. Starting from this observation, we propose a model of argumentation where the arguments are related to the sources and their acceptability is computed on the basis of the trustworthiness of the sources. Furthermore, our approach goes beyond this observation by providing a feedback such that the final quality of the arguments influences the source evaluation as well.

Stranders et al. [5] propose an approach to trust based on argumentation that aims at exposing the rationale behind such trusting decisions. The aim of our work is different: we are interested in evaluating the arguments proposed by the sources with respect to their trustworthiness, instead of explaining, thanks to argumentation theory, the decisions about trusting or not another agent.

Prade [4] presents a bipolar qualitative argumentative modeling of trust where trust and distrust are assessed independently. The author introduces also a notion of reputation which is viewed as an input information used by an agent for revising or updating his trust evaluation. Reputation contributes to provide direct arguments in favor or against a trust evaluation. In this paper, we do not use observed behavior and reputation to compute the trust value, and we model the socio-cognitive dynamics of trust such as the feedback and the trust dimensions, differently from [4].

Matt et al. [6] propose to construct a Dempster-Shafer belief function both from statistical data and from arguments in the context of contracts. We do not have arguments expressing the trustworthiness degree assigned to the other agents, but we accept the arguments depending on the trustworthiness of their sources. Moreover, in our model, the trustworthiness assigned to the arguments feeds back

to the sources dynamically changing their own trustworthiness. We distinguish also the two dimensions of sincerity and competence.

Tang et al. [23] and Parsons et al. [7] present a framework to introduce the sources in argumentation and to express the degrees of trust. They define trust-extended argumentation graphs in which each premise, inference rule, and conclusion is associated to the trustworthiness degree of the source proposing it. Thus, given two arguments rebutting each other, the argument whose conclusion has a higher trust value is accepted. They do not have the possibility to directly attack the trustworthiness of the sources as well as the trustworthiness of single arguments and attacks. Again, the feedback towards the source as well as the distinction between competence and sincerity is not considered. We do not express the degrees of trust in a fine-grained way as done in [23, 7].

da Costa Pereira et al. [9] propose a framework where argumentation theory is used in belief revision. In this framework, the arguments are weighted on the basis of the trustworthiness degree of the sources proposing them. The acceptability of the arguments is then computed by a labelling algorithm which assigns the arguments a fuzzy value, differently from Dung-like frameworks where arguments are either accepted or rejected. In this paper, we do not assign a numerical trust value to the sources, and we stay close to Dung-like frameworks. The framework of da Costa Pereira and colleagues has a number of limitations, such as the absence of a feedback mechanism, and the fact that the notion of trust is considered as a monolithic and not multidimensional concept.

A huge amount of research has been conducted on trust, and some of these works are described below, even if in this paper we limit our attention to the cognitive trust model of Castelfranchi and Falcone [1].

Castelfranchi and Falcone [1] stress the importance of this explicit cognitive account for trust in three ways. First, they criticize the game-theoretic view of trust which is prisoner of the Prisoner Dilemma mental frame, and reduce trust simply to a probability or perceived risk in decisions. Second, they find the quantitative aspects of trust (its strength or degree) on those mental ingredients (beliefs and goals) and on their strength. Third, they claim that this cognitive analysis of trust is fundamental for distinguishing between internal and external attribution which predict very different strategies for building or increasing trust; for founding mechanisms of image, reputation, persuasion, and argumentation in trust building. Apart from the cognitive model of Castelfranchi and Falcone [1] that define trust as "*a mental state, a complex attitude of an agent x towards another agent y about the behaviour/action a relevant for the goal g*", many other definitions have been provided in the literature.

In sociology, Gambetta [24] states that "*trust is the subjective probability by which an individual A expects that another individual B performs a given action on which its welfare depends*". Castelfranchi and Falcone [1] observe that this definition is correct. However, it is also quite a poor definition, since it just refers to one dimension of trust, i.e., predictability, while ignoring the "competence" dimension; it does not account for the meaning of "I trust B" where there is also the decision and the act of relying on B; and it does not explain what is such an evaluation made of and based on. Common elements of these definitions are a consistent degree of uncertainty and conflicting information associated with trust.

Another approach to model trust using modal logic is proposed by Lorini and Demolombe [25] where they present a concept of trust that integrates the truster's goal, the trustee's action ensuring the achievement of the truster's goal, and the trustee's ability and intention to do this action. In this paper, we do not refer to the actions of the sources, but we provide a model for representing the conflicts the sources have to deal with trust. The introduction of the actions in our cognitive model is left as future work, and it will allow also to model willingness (source s should think that source p not only is able and can do that action/task, but p actually will do what s needs). In this paper, we model only the competence and sincerity mental states of trust.

Another proposal is presented by Liau [26], in which the influence of trust on the assimilation of information into the source's mind is considered. The idea is that "if agent i believes that agent j has told him the truth on p, and he trusts the judgement of j on p, then he will also believe p". Extending the model by introducing goals to model the presented definitions is left for future work.

Wang and Singh [27], instead, understand trust in terms of belief and certainty: A's trust in B is reflected in the strength of A's belief that B is trustworthy. They formulate certainty in terms of evidence based on a statistical measure defined over a probability distribution of positive outcomes. Both Liau [26] and Wang and Singh [27] capture intuitions that play a role also in our approach, but they propose a simplified model of the nature and dynamics of trust, as opposed to the socio-cognitive model discussed in [1].

4. Modelling trust in meta-argumentation

In this section, we formally define our cognitive model of trust using meta-argumentation. Using the running example described in the introduction, we show how the model can be used to formally model it, and we present some desired properties of our model.

4.1. Information sources

The reason why abstract argumentation is not suited to model trust is that an argument, if it is not attacked by another acceptable argument, is considered acceptable. This prevents us from modeling the situation where, for an argument to be acceptable, it must be related to some trusted sources which provide the evidence for such an argument to be accepted. Without an explicit representation of the sources, it becomes impossible to talk about trust: the argument can only be attacked by conflicting information, but it cannot be made unacceptable due to the lack of trust in the source.

Modelling evidence is another challenge: sources are a particular type of evidence. Arguments needing evidence are well known in legal argumentation, where the notion of burden of proof has been introduced [12]. Meta-argumentation provides a means to model burden of proof in abstract argumentation without extending argumentation. The idea is to associate to each argument $a \in A$ put on the table, which is represented by means of meta-argument $acc(a)$, an auxiliary argument $W_{acc(a)}$ attacking it. Being auxiliary this argument is filtered out during the unflattening process. This means that without further information, just as being put on the table, argument a is not acceptable since it is attacked by the acceptable argument $W_{acc(a)}$, and there is no evidence defending it against this "default" attack, as visualized in Figure 2 for arguments a and b. This evidence is modeled by means of the attacks towards these auxiliary arguments, e.g., $W_{acc(a)}$, leading to a reinstatement of meta-argument $acc(a)$. Attacks are modeled as arguments as well, so they need evidence to be acceptable. For each auxiliary argument $Y_{a,b}$, representing the activation of the attack, we associate an auxiliary argument $W_{Y_{a,b}}$.

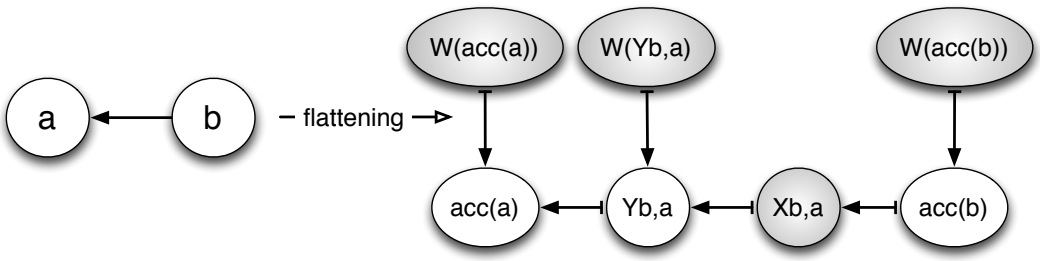

Figure 2: Arguments and attacks without evidence.

Sources are introduced in the meta-argumentation framework under the form of meta-arguments "*source s is trustable*", $trust(s)$, for all the sources s. Each argument a in the sources' mind is supported by means of an attack on $W_{acc(a)}$.

We represent the fact that one or more information sources support the same argument by letting them attack the same $W_{acc(a)}$ auxiliary argument. An example of multiple evidence is depicted in Figure 3. In the figures, we represent the meta-arguments associated to the information sources as boxes, and the arguments as circles where grey elements are the acceptable ones. As for arguments, an attack to become active needs some trusted agent.

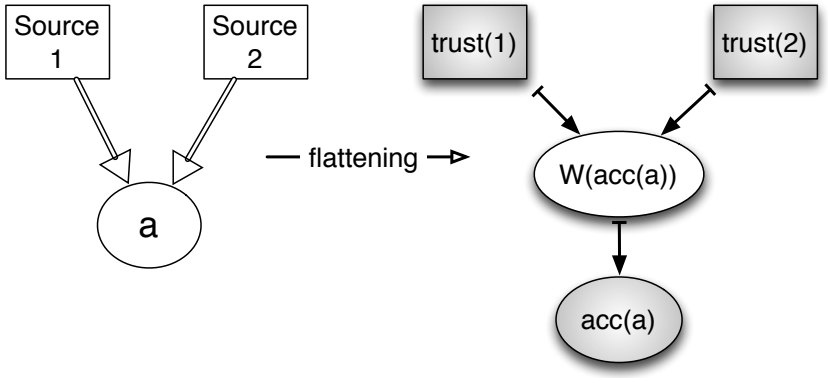

Figure 3: An example of multiple evidence.

Notice that the assumption that there must be evidence for an argument to be accepted is a very general and often used reasoning pattern, e.g., in causal reasoning, where everything needs to be explained, i.e., to have a cause / to be caused, as in the Yale shooting problem for instance. For more details about causal reasoning, see Bochman [28].

We have now to discuss which semantics we adopt for assessing the acceptability of the arguments and the sources. For example, suppose that two sources claim they are each untrustworthy. What is the extension? We adopt admissibility based semantics. We do not ask for completeness because if one wants to know whether a particular argument is acceptable, the whole model is not needed, just the part related to this particular argument is needed.

The reader should not be confused by the similarity between evidence and support [29]. The meaning of Boella et al. [29]'s notion of support is that if a is acceptable then b is acceptable too. Note that the supported argument b is acceptable (if not attacked) even without the support of a, i.e., a is not acceptable. Support exploits an auxiliary argument Z, but with some difference with the auxiliary argument W. First, given a supporting b, there is a $Z_{a,b}$ such that b attacks $Z_{a,b}$ and $Z_{a,b}$ attacks a, while, here, $W_{acc(a)}$ attacks the argument needing evidence. Sec-

ond, there is a Z meta-argument for each supporting argument, while, here, there is only one W meta-argument attacked by all the arguments and agents providing an evidence. For more details about our model of support in argumentation, see Boella et al. [29].

We extend the definition of EAF (Definition 6) by adding evidence provided by the information sources and second-order attacks, such as attacks from an argument or attack to another attack. For more details about second-order attacks in meta-argumentation, see [14, 15]. The unflattening function g and the acceptance function $\mathscr{E}'$ are defined as above.

Definition 8. *A trust-based extended argumentation framework $TEAF^2$ with second-order attacks is a tuple $\langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2 \rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2 \rangle, \rightarrow \rangle$ where for each source $1 \leq i \leq n$, $A_i \subseteq A \subseteq \mathscr{U}$ is a set of arguments, $\rightarrow \subseteq A \times A$, $\rightarrow_i$ is a binary relation on $A_i \times A_i$, $\rightarrow_i^2$ is a binary relation on $(A_i \cup \rightarrow_i) \times \rightarrow_i$.*

Definition 9 presents the instantiation of a $TEAF^2$ with second-order attacks as a set of partial frameworks of the sources using meta-argumentation.

Definition 9. *Given a $TEAF^2 = \langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2 \rangle \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2 \rangle, \rightarrow \rangle$, the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in A_1 \cup \ldots \cup A_n\} \cup \{W_{acc(a)} \mid a \in A_1 \cup \ldots \cup A_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that:*

- $acc(a) \longmapsto X_{a,b}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and* $X_{a,b} \longmapsto Y_{a,b}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and* $Y_{a,b} \longmapsto acc(b)$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and*
- $trust(i) \longmapsto W_{acc(a)}$ *iff* $a \in A_i$, *and* $W_{acc(a)} \longmapsto acc(a)$ *iff* $a \in A$, *and*
- $trust(i) \longmapsto W_{Y_{a,b}}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and* $W_{Y_{a,b}} \longmapsto Y_{a,b}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and*
- $acc(a) \longmapsto X_{a,b \rightarrow c}$ *iff* $a,b,c \in A_i$ *and* $a \rightarrow_i^2 (b \rightarrow_i c)$, *and* $X_{a,b \rightarrow c} \longmapsto Y_{a,b \rightarrow c}$ *iff* $a,b,c \in A_i$ *and* $a \rightarrow_i^2 (b \rightarrow_i c)$, *and* $Y_{a,b \rightarrow c} \longmapsto Y_{b,c}$ *iff* $a,b,c \in A_i$ *and* $a \rightarrow_i^2 (b \rightarrow_i c)$, *and*
- $Y_{a,b} \longmapsto Y_{c,d}$ *iff* $a,b,c \in A_i$ *and* $(a \rightarrow_i b) \rightarrow_i^2 (c \rightarrow_i d)$.

We say that source i is trustworthy when meta-argument $trust(i)$ is acceptable, and we say that i provides evidence in support of argument a (of the attack $a \rightarrow b$) when $a \in A_i$ (when $a \rightarrow b \in \rightarrow_i$), and $trust(i) \longmapsto W_{acc(a)}$ ($trust(i) \longmapsto W_{Y_{a,b}}$).

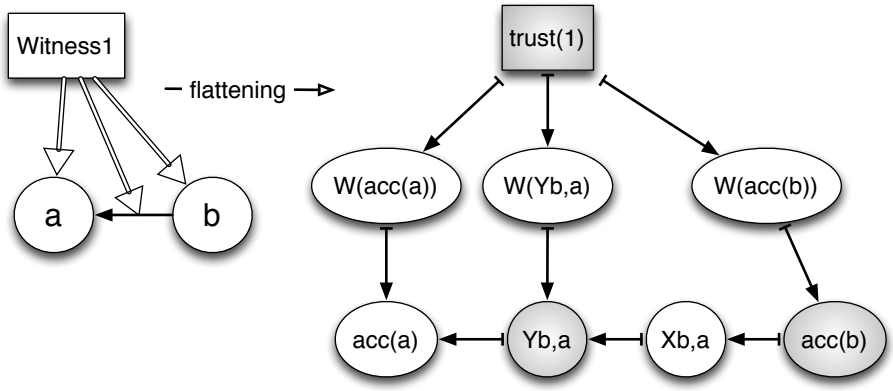

Figure 4: Introducing the sources in the argumentation frameworks.

Example 1. *Consider the informal dialogue provided in the introduction. We represent the sources in the argumentation framework, as shown in Figure 4. Witness1 proposes a and b and the attack $a \rightarrow b$. Using the flattening function of Definition 9, we add meta-argument $trust(1)$ for representing Witness1 in the framework, and we add meta-arguments $acc(a)$ and $acc(b)$ for the arguments of Witness1. Witness1 provides evidence for these arguments, and the attack $b \rightarrow a$ by attacking the respective auxiliary arguments W. In the remainder of the paper, we model the other conflicts highlighted in the dialogue.*

Let $trust(i)$ be the information source i and $acc(a)$ and $Y_{a,b}$ the argument a_i and the attack $a \rightarrow_i b$ respectively, as defined in Definitions 6 and 7. Meta-argument $trust(i)$ can provide evidence for $acc(a)$ and $Y_{a,b}$. Sources can attack other sources as well as their arguments and attacks. With a slight abuse of notation, we write $a \in \mathscr{E}'(EAF)$, even if the latter is a set of extensions, with the intended meaning that a is in some of the extensions of $\mathscr{E}'$. We now provide some properties of our model.

Proposition 1. *Assume admissibility based semantics, if an argument $a \in A$ is not supported by evidence, i.e., $a \notin A_i$ for all i, then a is not accepted, $a \notin \mathscr{E}'(EAF)$.*

Proof. We prove the contrapositive: if argument a is accepted, then argument a is supported by evidence. Assume argument a is accepted. Then auxiliary argument $W_{acc(a)}$ is rejected due to the conflict-free principle. Meta-argument $acc(a)$ is defended, so $W_{acc(a)}$ is attacked by an accepted argument using admissible semantics. Auxiliary argument $W_{acc(a)}$ can only be attacked by meta-argument $trust(i)$. We conclude that a is supported by evidence. □

Proposition 1 is strengthened to Proposition 2.

Proposition 2. *If an argument a is not supported, $a \notin A_i$, then the extensions $\mathscr{E}'(EAF)$ are precisely the same as the extensions of the $AF = \langle A, \rightarrow \rangle$ in which $a \notin A$, and the attacks on a or from a do not exist, i.e., $b \rightarrow a \notin \rightarrow$ and $a \rightarrow c \notin \rightarrow$.*

Proof. Assume argument a is not supported by evidence. This means that meta-argument $W_{acc(a)}$ is accepted, and meta-argument $acc(a)$ is not accepted. Assume there exist an argument b such that b attacks a, $b \rightarrow a$, and an argument c such that a attacks c, $a \rightarrow c$. We prove that the extensions of the EAF with argument a are precisely the same as the extensions of the AF in which a does not exist, and the attacks $b \rightarrow a$ and $a \rightarrow c$ do not exist either. We use case analysis.

Case 1 Assume arguments b and c are not attacked, or they are attacked by unaccepted arguments. Then, we have that meta-argument $acc(b)$ is accepted, meta-argument $Y_{b,a}$ is accepted and meta-argument $acc(a)$ is not accepted as assumed, and meta-argument $acc(c)$ is accepted, meta-argument $X_{a,c}$ is accepted, $Y_{a,c}$ is not accepted and $acc(a)$ is not accepted as assumed. The extension of this EAF includes b and c, but it does not include a.

Case 2 Assume arguments b and c are attacked by accepted arguments. Then, we have that meta-argument $acc(b)$ is unaccepted, meta-argument $Y_{b,a}$ is unaccepted and meta-argument $acc(a)$ is not accepted as assumed, and meta-argument $acc(c)$ is unaccepted, meta-argument $X_{a,c}$ is accepted, $Y_{a,c}$ is not accepted and $acc(a)$ is not accepted as assumed. The extension of this EAF does not include a, b, and c.

Case 3 Assume argument b is not attacked or it is attacked by unaccepted arguments, and c is attacked by accepted arguments. Then, we have that meta-argument $acc(b)$ is accepted, meta-argument $Y_{b,a}$ is accepted and meta-argument $acc(a)$ is not accepted as assumed, and meta-argument $acc(c)$ is unaccepted, meta-argument $X_{a,c}$ is accepted, $Y_{a,c}$ is not accepted and $acc(a)$ is not accepted as assumed. The extension of this EAF includes b, but it does not include a and c.

Case 4 Assume argument b is attacked by accepted arguments and c is not attacked or it is attacked by unaccepted arguments. Then, we have that meta-argument $acc(b)$ is unaccepted, meta-argument $Y_{b,a}$ is unaccepted and meta-argument $acc(a)$ is not accepted as assumed, and meta-argument $acc(c)$ is accepted, meta-argument $X_{a,c}$ is accepted, $Y_{a,c}$ is not accepted and $acc(a)$ is

not accepted as assumed. The extension of this EAF includes c, but it does not include a and b.

Now we consider the same EAF without argument a, such that the attacks $b \rightarrow a$ and $a \rightarrow c$ do not exist.

Case 1 Assume arguments b and c are not attacked, or they are attacked by unaccepted arguments. Then, we have that meta-argument $acc(b)$ is accepted and meta-argument $acc(c)$ is accepted too. Each extension of this AF includes b and c.

Case 2 Assume arguments b and c are attacked by accepted arguments. Then, we have that meta-argument $acc(b)$ is unaccepted, and meta-argument $acc(c)$ is unaccepted either. Each extension of this AF does not include b, and c.

Case 3 Assume argument b is not attacked or it is attacked by unaccepted arguments, and c is attacked by accepted arguments. Then, we have that meta-argument $acc(b)$ is accepted, and meta-argument $acc(c)$ is unaccepted. Each extension of this AF includes b, but it does not include c.

Case 4 Assume argument b is attacked by accepted arguments and c is not attacked or it is attacked by unaccepted arguments. Then, we have that meta-argument $acc(b)$ is unaccepted, and meta-argument $acc(c)$ is accepted. Each extension of this AF includes c, but it does not include b.

Thus, the extensions of the EAF including argument a without evidence, and the EAF not including argument a are the same. □

Proposition 3. *If an attack $a \rightarrow b$ is not supported, i.e., $a \rightarrow b \notin \rightarrow_i$, then the extensions $\mathscr{E}'(EAF)$ are precisely the same as the extensions of the $AF = \langle A, \rightarrow \rangle$, in which the attack does not exist, $a \rightarrow b \notin \rightarrow$.*

The proof of Proposition 3 follows Proof 4.1.

Proposition 4. *Assume EAF is a framework in which argument a is supported by the trustworthy source i, and there is another trustworthy source j. In that case, the extensions are the same if also j provides an evidence in support of a.*

Proof. Assume argument a is supported by the trustworthy source i. This means that $trust(i)$ is accepted. It supports by evidence argument a which means that meta-argument $trust(i)$ attacks meta-argument $W_{acc(a)}$: meta-argument $trust(i)$ is

accepted, thus meta-argument $X_{trust(i),W_{acc(a)}}$ is unaccepted, meta-argument $Y_{trust(i),W_{acc(a)}}$ is accepted and meta-argument $W_{acc(a)}$ is unaccepted. Thus, meta-argument $acc(a)$ is accepted. We use case analysis.

Case 1 : Let argument a not be attacked or be attacked by unaccepted arguments. This means that meta-argument $acc(a)$ is accepted, and argument a is part of each extension of the EAF.

Case 2 : Let argument a be attacked by accepted arguments. This means that meta-argument $acc(a)$ is not accepted, and argument a is not part of the extensions of the EAF.

Assume there is another trustworthy source j. This means that meta-argument $trust(j)$ is accepted. This source supports by evidence argument a, too. This means that $trust(j)$ attacks meta-argument $W_{acc(a)}$: meta-argument $trust(j)$ is accepted, thus meta-argument $X_{trust(j),W_{acc(a)}}$ is unaccepted, meta-argument $Y_{trust(j),W_{acc(a)}}$ is accepted and meta-argument $W_{acc(a)}$ is unaccepted. Thus, meta-argument $acc(a)$ is accepted. We use case analysis.

Case 1 : Let argument a not be attacked or be attacked by unaccepted arguments. This means that meta-argument $acc(a)$ is accepted, and argument a is part of each extension of the EAF.

Case 2 : Let argument a be attacked by accepted arguments. This means that meta-argument $acc(a)$ is not accepted, and argument a is not part of the extensions of the EAF.

Thus, the extensions of the EAF are the same if there is also another source j, in addition to i, supporting argument a. □

4.2. Evidence for arguments

The evidence in favor of the arguments is an evidence provided by the agents for the arguments/attacks they propose. At the meta-level, this is modeled as an attack from meta-argument $trust(i)$ to W auxiliary arguments. However, there are other cases in which more evidence is necessary to support the acceptability of an argument. Consider the case of Witness1. His trustworthiness is attacked by Witness2. What happens to the evidence provided by Witness1? Since the source is not trustworthy then it cannot provide evidence. Meta-argument $trust(1)$ becomes not acceptable and the same happens to all its arguments and attacks. What is needed to make them acceptable again is more evidence. This evidence can be

provided under the form of another argument which reinstates the acceptability of these information items.

Definition 9 allows only the sources to directly provide evidence for the information items. As for Witness5 and Witness6 in the dialogue, sources can provide evidence also by means of other arguments. This cannot be represented using the extended argumentation framework of Definition 9, this is why we need to extend it with an evidence relation $\not\rightarrow$ representing evidence provided under the form of arguments for the information items of the other sources.

Definition 10. *A $TEAF^2$ with evidence is a tuple $\langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2, \not\rightarrow_1 \rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2, \not\rightarrow_n \rangle, \rightarrow \rangle$ where $\not\rightarrow_i$ is a binary relation on $A_i \times A_j$ and the set of meta-arguments MA is $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in A_1 \cup \ldots \cup A_n\} \cup \{W_{acc(a)} \mid a \in A_1 \cup \ldots \cup A_n\}$ and $\longmapsto \subseteq MA \times MA$ is a binary relation on MA such that hold the conditions of Definition 9, and: $acc(a) \longmapsto W_{acc(b)}$ iff $a \in A_i, b \in A_j$ and $a \not\rightarrow_i b$, and $W_{acc(b)} \longmapsto acc(b)$ iff $a \in A_i, b \in A_j$ and $a \not\rightarrow_i b$.*

We say that a source i supports the evidence provided by other source j to argument a when $a \in A_i, b \in A_j$, and $acc(a) \longmapsto W_{acc(b)}$.

The following properties hold for Definition 10.

Proposition 5. *If there are multiple arguments $a_1 \in A_1, \ldots, a_n \in A_n$ providing evidence for an argument $b \in A_k$ (or an attack), and there are no attacks on the arguments, $c_1 \rightarrow a_1 \not\in \rightarrow_1, \ldots, c_n \rightarrow a_n \not\in \rightarrow_n$, then b (or the attack) is accepted, $b \in \mathscr{E}'(EAF)$, iff at least one of the sources providing evidence for arguments $a_1, \ldots, a_n$ is trustworthy, i.e., $trust(j) \in \mathscr{E}(f(EAF))$ with $j \in 1, \ldots, n$.*

Proof. Assume argument b is not directly supported by evidence by an information source or the source supporting it is untrustworthy. This means that meta-argument $W_{acc(b)}$ is accepted, and meta-argument $acc(b)$ is not accepted. Assume now that argument b is not attacked by other arguments, or it is attacked by unaccepted arguments, and assume there are n arguments $a_1, \ldots, a_n$ providing evidence for argument b. Assume there not exist argument c_i such that it attacks a_i, and c_i is accepted.

First, we show that if there is at least one trustworthy source proposing an argument a_i which provides evidence for argument b, then b is accepted. This means that $trust(i)$ is accepted for $1 \leq i \geq n$. Then $W_{acc(a_i)}$ is unaccepted, and $acc(a)$ is accepted, $W_{acc(b)}$ is unaccepted and $acc(b)$ is accepted.

Now, we show that if argument b is accepted then there is at least one trustworthy source providing evidence for it through argument a_i. This means that $acc(b)$

is accepted, and $W_{acc(b)}$ is not accepted. Thus there is at least on $Y_{acc(a_i),W_{acc(b)}}$ which is accepted. This means that $acc(a_i)$ is accepted, and $trust(i)$ is accepted. □

Proposition 6. *Suppose two sources i and j provide evidence through arguments b and c respectively for the same argument a, i.e., $b \leadsto a \in \leadsto_i$ and $c \leadsto a \in \leadsto_j$, then it is the same whether a trustworthy source k supports the evidence provided by i or j, i.e., $d \in A_k$.*

Proof. Assume source k is trustworthy. Source k provides evidence for argument d. Assume there are no other attacks on d. This means that meta-argument $trust(k)$ attacks meta-argument $W_{acc(d)}$ and $W_{acc(d)}$ attacks meta-argument $acc(d)$. The accepted meta-arguments are $acc(d)$ and $trust(k)$. We use case analysis.

Case 1 : Let the sources i and j be trustworthy, and let their arguments not be attacked by other arguments. This means that meta-arguments $trust(i)$ and $trust(j)$ are accepted, meta-arguments $acc(b)$ and $acc(c)$ are accepted and meta-argument $acc(a)$ is accepted. The support of source k through argument d consists in an attack from meta-argument $acc(d)$ to meta-argument $W_{acc(b)}$ or to meta-argument $W_{acc(c)}$. Both these meta-arguments are not accepted because of the attacks from $trust(i)$ and $trust(j)$, respectively.

Case 2 : Let the sources i and j be untrustworthy, and let their arguments not be attacked by other arguments. This means that meta-arguments $trust(i)$ and $trust(j)$ are unaccepted. Thus, meta-arguments $acc(b)$ and $acc(c)$ are unaccepted. The evidence provided through argument d by source k consists in an attack from meta-argument $acc(d)$ to meta-argument $W_{acc(b)}$ or $W_{acc(c)}$. Independently on which meta-argument is attacked, this means that meta-argument $acc(b)$ or meta-argument $acc(c)$ is accepted, meta-argument $W_{acc(a)}$ is not accepted and meta-argument $acc(a)$ is accepted.

Case 3 : Let source i (or j) be trustworthy and source j (or i) be untrustworthy, and let their arguments not be attacked by other arguments. This means that meta-argument $trust(i)$ is accepted and meta-argument $trust(j)$ is not trustworthy, meta-argument $W_{acc(b)}$ is unaccepted and meta-argument $W_{acc(c)}$ is accepted, meta-argument $acc(b)$ is accepted and meta-argument $acc(c)$ is not accepted. Thus meta-argument $W_{acc(a)}$ is not accepted and meta-argument $acc(a)$ is accepted. The evidence provided through argument d to argument a does not change if meta-argument $acc(c)$ attacks meta-argument

$W_{acc(b)}$ or $W_{acc(c)}$, because meta-argument $W_{acc(a)}$ is attacked by both $acc(b)$ and $acc(c)$. Thus, meta-argument $acc(a)$ is accepted independently from the evidence provided by argument d.

□

Example 2. *Consider the dialogue in the introduction. Argument g by Witness6 is an evidence for argument f by Witness5. This evidence is expressed in meta-argumentation in the same way as evidence provided by the sources, such as an attack to $W_{acc(f)}$ attacking $acc(f)$. In this case, it is meta-argument $acc(g)$ which attacks $W_{acc(f)}$, as visualized in Figure 5.*

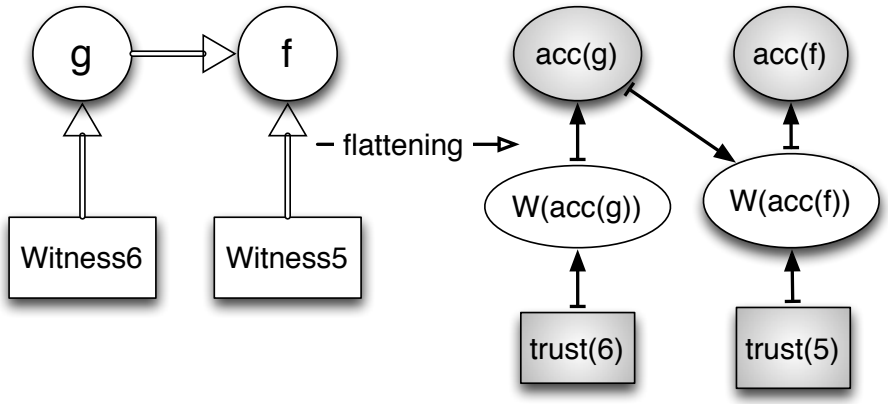

Figure 5: Introducing evidence for the arguments.

4.3. Focused trust relationships

In our model, trust is represented as the absence of an attack towards the sources or towards their information items, and as the presence of evidence in favor of the pieces of information. On the contrary, the distrust relationship is modeled as a lack of evidence in support of the information items or as a direct attack towards the sources and their pieces of information.

In the informal dialogue, Witness2 attacks the trustworthiness of Witness1 as a credible witness. In this way, she is attacking each argument and attack proposed by Witness1. Witness4, instead, is not arguing against Witness3 but she is arguing against the attack $d \rightarrow b$ as it is proposed by Witness3. Finally, for Witness2 the untrustworthiness of Witness6 is related only to the argument g. We propose a

focused view of trust in which the information sources may be attacked for being untrustworthy or for being untrustworthy only concerning a particular argument or attack. Definition 11 presents an *EAF* in which a new relation *DT* between sources is given to represent distrust.

Definition 11. *A trust-based extended argumentation framework* $DTEAF^2$ *is a tuple* $\langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2, \not\rightarrow_1, DT_1\rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2, \not\rightarrow_n, DT_n\rangle, \rightarrow\rangle$ *where for each source* $1 \leq i \leq n$, $A_i \subseteq A \subseteq \mathscr{U}$ *is a set of arguments,* $\rightarrow \subseteq A \times A$, $\rightarrow_i \subseteq A_i \times A_i$ *is a binary relation,* $\rightarrow_i^2$ *is a binary relation on* $(A_i \cup \rightarrow_i) \times \rightarrow_i$, $\not\rightarrow_i$ *is a binary relation on* $A_i \times A_j$, *and* $DT \subseteq A_i \times \vartheta$ *is a binary relation such that* $\vartheta = j$ *or* $\vartheta \in A_j$ *or* $\vartheta \in \rightarrow_j$.

Definition 12 shows how to instantiate a $DTEAF^2$ enriched with a distrust relation with meta-arguments. In particular, the last three points of Definition 12 model, respectively, a distrust relationship towards an agent, a distrust relationship towards an argument, and a distrust relationship towards an attack. The unflattening function g and the acceptance function $\mathscr{E}'$ are defined as above.

Definition 12. *Given a* $DTEAF^2 = \langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2, \not\rightarrow_1, DT_1\rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2, \not\rightarrow_n, DT_n\rangle, \rightarrow\rangle$, *see Definition 11, the set of meta-arguments MA is* $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in A_1 \cup \ldots \cup A_n\} \cup \{W_{acc(a)} \mid a \in A_1 \cup \ldots \cup A_n\}$ *and* $\longmapsto \subseteq MA \times MA$ *is a binary relation on MA such that hold the conditions of Definitions 9 and 10, and:*

- $acc(a) \longmapsto X_{a,b}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and* $X_{a,b} \longmapsto Y_{a,b}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and* $Y_{a,b} \longmapsto acc(b)$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and*

- $trust(i) \longmapsto X_{trust(i),W_{acc(a)}}$ *iff* $a \in A_i$, *and* $X_{trust(i),W_{acc(a)}} \longmapsto Y_{trust(i),W_{acc(a)}}$ *iff* $a \in A_i$, *and* $Y_{trust(i),W_{acc(a)}} \longmapsto W_{acc(a)}$ *iff* $a \in A_i$, *and* $W_{acc(a)} \longmapsto acc(a)$ *iff* $a \in A_i$, *and*

- $trust(i) \longmapsto X_{trust(i),W_{Y_{a,b}}}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and* $X_{trust(i),W_{Y_{a,b}}} \longmapsto Y_{trust(i),W_{Y_{a,b}}}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and* $Y_{trust(i),W_{Y_{a,b}}} \longmapsto W_{Y_{a,b}}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and* $W_{Y_{a,b}} \longmapsto Y_{a,b}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$, *and*

- $trust(i) \longmapsto W_{acc(a)}$ *iff* $a \in A_i$ *and* $aDT_i trust(j)$, *and* $W_{acc(a)} \longmapsto acc(a)$ *iff* $a \in A$ *and* $aDT_i trust(j)$, *and* $acc(a) \longmapsto X_{acc(a),trust(j)}$ *iff* $a \in A_i$ *and* $aDT_i trust(j)$, *and* $X_{acc(a),trust(j)} \longmapsto Y_{acc(a),trust(j)}$ *iff* $a \in A_i$ *and* $aDT_i trust(j)$, *and* $Y_{acc(a),trust(j)} \longmapsto trust(j)$ *iff* $a \in A_i$ *and* $aDT_i trust(j)$, *and*

- $trust(i) \longmapsto W_{acc(a)}$ *iff* $a \in A_i, b \in A_j$ *and* aDT_ib, *and* $W_{acc(a)} \longmapsto acc(a)$ *iff* $a \in A, b \in A_j$ *and* aDT_ib, *and* $acc(a) \longmapsto X_{acc(a),Y_{trust(j),W_{acc(b)}}}$ *iff* $a \in A_i, b \in A_j$ *and* aDT_ib, *and* $X_{acc(a),Y_{trust(j),W_{acc(b)}}} \longmapsto Y_{acc(a),Y_{trust(j),W_{acc(b)}}}$ *iff* $a \in A_i, b \in A_j$ *and* aDT_ib, *and* $Y_{acc(a),Y_{trust(j),W_{acc(b)}}} \longmapsto Y_{trust(j),W_{acc(b)}}$ *iff* $a \in A_i, b \in A_j$ *and* aDT_ib, *and*

- $trust(i) \longmapsto W_{acc(a)}$ *iff* $a \in A_i, b, c \in A_j$ *and* $aDT_i(b \rightarrow_j c)$, *and* $W_{acc(a)} \longmapsto acc(a)$ *iff* $a \in A, b, c \in A_j$ *and* $aDT_i(b \rightarrow_j c)$, *and* $acc(a) \longmapsto X_{acc(a),Y_{trust(j),W_{Y_{b,c}}}}$ *iff* $a \in A_i, b, c \in A_j$ *and* $aDT_i(b \rightarrow_j c)$, *and* $X_{acc(a),Y_{trust(j),W_{Y_{b,c}}}} \longmapsto Y_{acc(a),Y_{trust(j),W_{Y_{b,c}}}}$ *iff* $a \in A_i, b, c \in A_j$ *and* $aDT_i(b \rightarrow_j c)$, *and* $Y_{acc(a),Y_{trust(j),W_{Y_{b,c}}}} \longmapsto Y_{trust(j),W_{Y_{b,c}}}$ *iff* $a \in A_i, b, c \in A_j$ *and* $aDT_i(b \rightarrow_j c)$.

We say that a source j is untrustworthy when there is an attack from an argument $a \in A_i$ to j, $aDT_itrust(j)$. We say that an argument $a \in A_j$ or attack $a \rightarrow_j b \in \rightarrow_j$ is untrustworthy when there is an attack from an argument $c \in A_i$ to a or $a \rightarrow_j b$, cDT_ia or $cDT_i(a \rightarrow_j b)$.

Proposition 7. *Assume that source i is the only source providing evidence for argument $a \in A_i$ and attack $c \rightarrow b \in \rightarrow_i$, and assume admissibility based semantics. If the information source i is considered to be untrustworthy, then a and $c \rightarrow b$ are not acceptable.*

Proof. We prove the contrapositive: if the arguments and attacks supported by an information source i are acceptable then the information source i is considered to be trustworthy. Assume the source supports argument a and the attack $c \rightarrow b$ and assume that this argument and this attack are acceptable. Then auxiliary arguments $W_{acc(a)}$ and $W_{Y_{c,b}}$ are rejected due to the conflict-free principle. Meta-arguments $acc(a)$ and $Y_{c,b}$ are defended, thus $W_{acc(a)}$ and $W_{Y_{c,b}}$ are attacked by an acceptable argument, using admissible semantics. We assumed that this argument and this attack have no other evidence, so auxiliary arguments $W_{acc(a)}$ and $W_{Y_{c,b}}$ can only be attacked by meta-argument $trust(i)$. Since they are attacked by an acceptable argument, we conclude that the source i is acceptable. □

Example 3. *Figure 6.a shows that Witness2 attacks the trustworthiness of Witness1 by means of argument c. In meta-argumentation, we have that $trust(2)$ provides evidence for $acc(c)$ by attacking meta-argument $W_{acc(c)}$ and, with meta-arguments X, Y, it attacks $trust(1)$. This means that if Witness1 is untrustworthy*

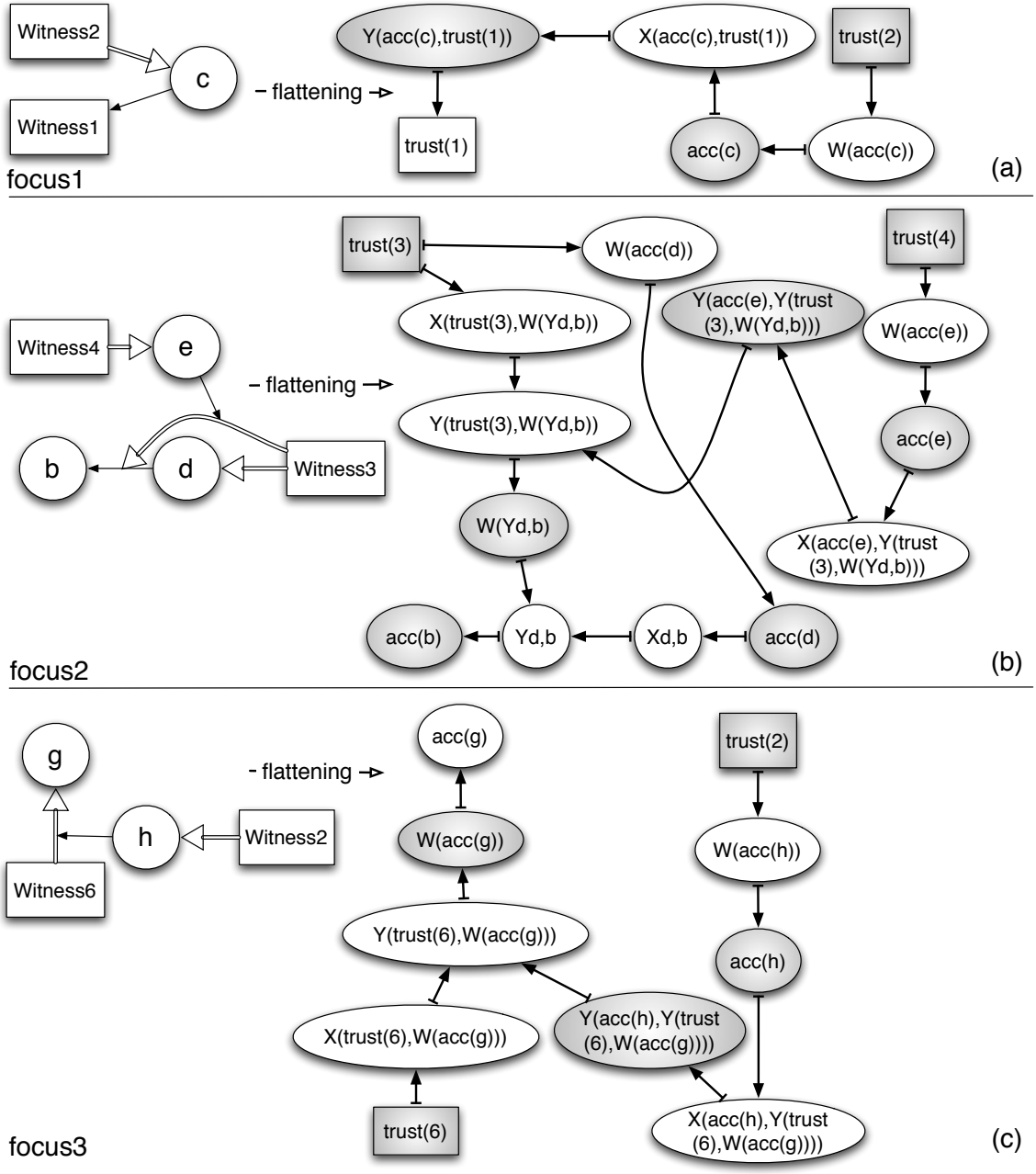

Figure 6: Focused trust in argumentation.

then each of his arguments and attacks cannot be acceptable either, if there is no more evidence. The set of acceptable arguments for the meta-argumentation framework is $\mathscr{E}(f(focus1)) = \{trust(2), acc(c), Y_{acc(c),trust(1)}\}$. *In Figure 6.b-c, instead, the attack is directed against a precise information item provided by the source. In particular, Witness4 attacks the attack* $d \rightarrow b$ *as provided by Witness3. This is achieved in meta-argumentation by means of an attack from meta-argument* $acc(e)$*, for which* $trust(4)$ *provides evidence, to the attack characterized by auxiliary argument* $Y_{d,b}$. *The set of acceptable arguments is* $\mathscr{E}(f(focus2)) = \{trust(4), trust(3), acc(d), acc(e), acc(b), Y_{acc(e),Y_{trust(3),W_{Y_{b,d}}}}, W_{Y_{d,b}}\}$. *Witness3's attack* $d \rightarrow b$ *is evaluated as untrustworthy by Witness4 and thus it is not acceptable. Finally, Witness2 evaluates Witness6 as untrustworthy concerning argument g. In meta-argumentation,* $trust(2)$*, by means of meta-argument* $acc(h)$*, attacks meta-argument* $acc(g)$ *proposed by* $trust(6)$. *The set of acceptable arguments is* $\mathscr{E}(f(focus3)) = \{trust(2), trust(6), acc(h), Y_{acc(h),Y_{trust(6),W_{acc(g)}}}, W_{acc(g)}\}$.

4.4. Feedback from information items to sources and back

In the previous sections, we have introduced the information sources in the argumentation framework in order to deal with the conflicts about trust. Moreover, in our framework, the agents are allowed to attack the trustworthiness of the other information sources or the trustworthiness of the single information items the sources propose. The relation, concerning trust, among the sources and the arguments or attacks they support is in one direction only. In particular, if an agent is considered not to be trustworthy, then also all the information items proposed by such an agent are considered untrustworthy. But what happens to the trustworthiness of an agent which is not directly attacked but it has all its information items (or at least *n* information items) attacked? In the current framework, these attacked items does not effect the trustworthiness of the sources proposing them, e.g., if a source has the trustworthiness of all its information items attacked, the source's trustworthiness is accepted.

The idea proposed by Castelfranchi and Falcone [1] is that there is a bidirectional link between the source and its information items: the provided data is more or less believable on the basis of the source's trustworthiness, but there is feedback such that the invalidation of the data feeds back on the sources' credibility. The overall amount and sign (increment or decrement) of the feedback depends on how much the overall quality of the message surprises the agent, with respect to its prior assessment of the source trustworthiness. This captures the principle that information quality should change one's assessment of its source only when

the agent learns something new about the capacity of the source to deliver information of either high or low quality. In other words, there should be a feedback on the source only when the quality of its argument tells me something new about the source's trustworthiness, revealing my previous opinion to be wrong. Otherwise, the quality of the new argument just confirms my previous assessment of the source, and confirmation, by definition, consolidates a pre-existing judgment, rather than modifying it. This points to the role of prediction in feedback dynamics from arguments to sources, and this prediction is based on the pre-existing degree of trustworthiness of the source of a given argument. In this paper, we does not represent the increment of the feedback towards the information source. In our framework, a trustworthy source is mirrored in an accepted meta-argument of the kind $trust(i)$, and this acceptability cannot be improved. The representation of this kind of feedback would be possible in numerical approaches to trust representation in argumentation, as proposed for instance by da Costa Pereira et al. [9] and Parsons et al. [7].

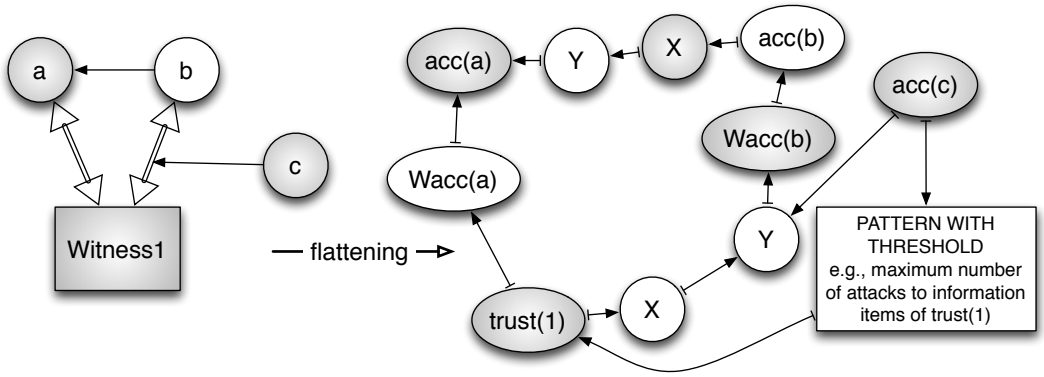

Figure 7: Feedback between the information items and sources.

In this section, we rely on this analysis of the trust dynamics phenomenon, in order to model the feedback from the information items to the sources. For instance, the fact that the major part of the arguments of a source are considered untrustworthy is seen as a negative experience, and leads to the decrease of the trustworthiness of the sources itself. In this paper, we do not consider the unpredictable cases analyzed by Castelfranchi and Falcone [1], where trust decreases with positive experiences, and increases with negative ones. The representation of these cases is left as future work.

We introduce the feedback from the information items to the sources, in such a way that, following different criteria, the untrustworthiness of the items influences

the trustworthiness of its source. The general idea of our approach is visualized in Figure 7. First, we insert in the framework a pattern which is activated if the number of attacks to this pattern exceeds a certain threshold. In this case, the pattern activates an attack towards the meta-argument representing the information source. The activation pattern is visualized in Figure 8, and it comes from the idea of conjunctive and proof standard patterns like those defined by Villata et al. [30]. The arguments attacking the information items proposed by the source attack also the pattern, in particular, each argument *arg* attacking the items attacks also one of the X meta-arguments of the pattern. These meta-arguments conjunctively attack argument s, which attacks the meta-argument representing the source. The pattern acts like a filter that raises the attack against the source only if the amount of incoming attacks is achieved.

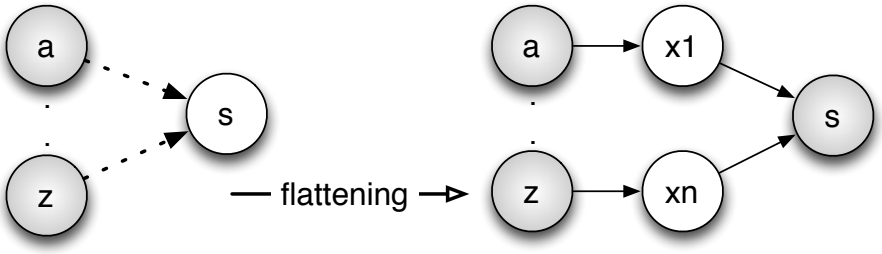

Figure 8: The activation pattern with a threshold of n arguments.

Second, for each attack to the trustworthiness of one of the information items of a source, this attack is duplicated and it is addressed also towards the pattern which attacks the information source. Summarizing, every attack to the arguments or attacks of a source is addressed also, i.e., towards a pattern which has the aim to attack directly the trustworthiness of the source, if the number of attack exceeds the given threshold.

Example 4. *Let us consider now the example proposed in Figure 7. In the informal dialogue, Witness1 proposes two arguments a and b, and the attack between them. Consider now the introduction of a new argument n, which attacks the trustworthiness of argument b as proposed by Witness1. In the flattened framework, the meta-argument* $trust(1)$ *provides evidence for meta-arguments* $acc(a)$ *and* $acc(b)$ *by attacking the auxiliary arguments* $W_{acc(a)}$ *and* $W_{acc(b)}$*. The attack of the new argument is addressed from meta-argument* $acc(n)$ *to the auxiliary argument* $Y_{W_{acc(b)}}$ *which attacks* $W_{acc(b)}$*. Since we are interested in modeling also the feedback from the information items to the sources, we add an additional attack*

from meta-argument $acc(n)$ to the pattern we use to measure the number of attacks to the information items proposed by Witness1. From this pattern, an attack is raised against the meta-argument $trust(1)$. If the number of attacks towards the pattern overcomes the given threshold, then the attack against $trust(1)$ becomes active, and $trust(1)$ becomes unacceptable, i.e., Witness1 is considered untrustworthy: so argument a is not acceptable.

We model feedback using the pattern associated with the threshold in order to maintain the choice of meta-argumentation, and avoiding the introduction of numerical techniques, as done for instance by da Costa Pereira et al. [9].

4.5. Modelling trust as a multidimensional concept

In this section, we investigate two dimensions of trust that have to be independently evaluated such as the sincerity or credibility of a source and its competence. We simplify Castelfranchi and Falcone's model [1], and focus only on two broad categories of relevant features in the source: competence (to what extent the source is deemed able to deliver a correct argument), and sincerity (to what extent the source is considered willing to provide a correct argument), both of which contribute to determine the source's overall trustworthiness. The evaluations of competence and sincerity are allowed to change across different domains. For instance, a reliable doctor will be considered competent in the health domain, but not necessarily so when suggesting a restaurant; conversely, a food critic is typically assumed to be trustworthy on the latter domain but not on the former. Similarly, one might think that a colleague who is competing with her for a promotion is likely to be insincere in giving her tips on how to improve her career, and yet there is no reason to doubt his sincerity when he suggests a movie. Here, we consider competence and sincerity as two possible dimensions for assessing the trustworthiness of a source.

We represent competence and sincerity using meta-arguments, and the attacks to these meta-arguments represent the conflicts about trust regarding the precise dimension of trust. The introduction in our framework of these two dimensions is visualized in Figure 9. We start from the usual situation in which an information source supports an argument, namely Witness7 supports argument *i* in the informal dialogue. We want to distinguish the two possible conflicts concerning argument *i*: a conflict meaning that Witness7 is considered untrustworthy on the competence regarding argument *i*, and a conflict meaning that Witness7 is considered untrustworthy on the sincerity in proposing argument *i*. An example of the first case is given in the dialogue by the attack of argument *l* to argument *i*, and

an example of the second case is given by the attack of argument m to argument i. Note that even if both arguments l and m attack argument i, they attack different dimensions of argument i.

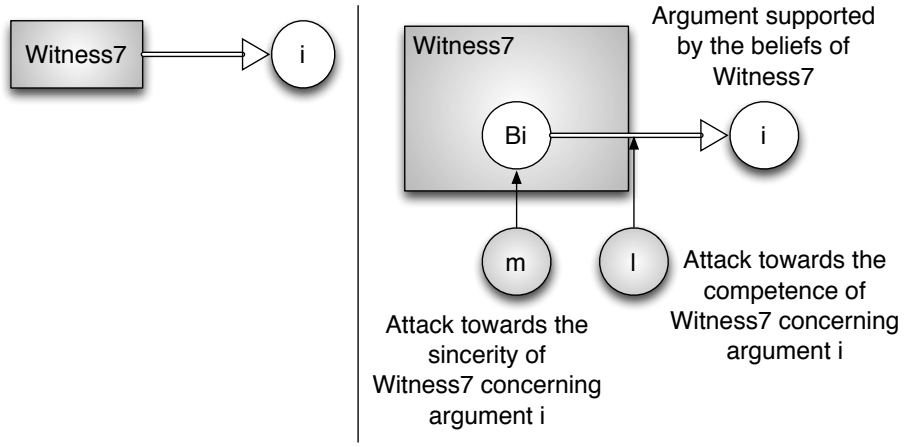

Figure 9: Modelling competence and sincerity.

We model sincerity and competence as visualized in Figure 9. Meta-argument B_i represents the belief associated to the information source concerning argument i, and it means "*the source believes argument i*" where argument i is the argument supported by the beliefs of the source. The meta-argument B_i supports argument i, as a result of the competence attributed to the source. In this framework, an attack towards the sincerity of the source is addressed against the meta-argument representing the belief of the source, i.e., against meta-argument B_i. An attack towards the competence of the source is addressed, instead, against the support relation between meta-argument B_i and argument i. This attack means that the source believes argument i but it is not evaluated competent concerning i. Note that an attack towards argument i is treated as in the previous sections, since it is a direct attack towards the content of argument i.

Definition 13. *A trust-based extended argumentation framework* $DTEAF^2_{CS}$ *is a tuple* $\langle A, \langle A_1, \rightarrow_1, \rightarrow_1^2, \not\rightarrow_1, DT_1, DT_{1s}, DT_{1c}\rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow_n^2, \not\rightarrow_n, DT_n, DT_{ns}, DT_{nc}\rangle, \rightarrow \rangle$ *where for each source* $1 \leq i \leq n$, $A_i \subseteq A \subseteq \mathscr{U}$ *is a set of arguments,* $\rightarrow \subseteq A \times A$, $\rightarrow_i \subseteq A_i \times A_i$ *is a binary relation,* $\rightarrow_i^2$ *is a binary relation on* $(A_i \cup \rightarrow_i) \times \rightarrow_i$, $\not\rightarrow_i$ *is a binary relation on* $A_i \times A_j$, *and* $DT \subseteq A_i \times \vartheta$ *is a binary relation such that* $\vartheta = j$, *and* $DT_s \subseteq A_i \times \vartheta$ *is a binary relation such that* $\vartheta \in A_j$ *or* $\vartheta \in \rightarrow_j$, *and* $DT_c \subseteq A_i \times \vartheta$ *is a binary relation such that* $\vartheta \in A_j$ *or* $\vartheta \in \rightarrow_j$.

Definition 14 shows how to instantiate an extended argumentation framework enriched with a distrust relation, which distinguishes distrust concerning competence and sincerity. The unflattening function g and the acceptance function $\mathscr{E}'$ are defined as above.

Definition 14. *Given a* $DTEAF^2_{CS} = \langle A, \langle A_1, \rightarrow_1, \rightarrow^2_1, \not\rightarrow_1, DT_1, DT_{1s}, DT_{1c}\rangle, \ldots, \langle A_n, \rightarrow_n, \rightarrow^2_n, \not\rightarrow_n, DT_n, DT_{ns}, DT_{nc}\rangle, \rightarrow\rangle$*, see Definition 13, the set of meta-arguments MA is* $\{trust(i) \mid 1 \leq i \leq n\} \cup \{acc(a) \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{X_{a,b}, Y_{a,b} \mid a,b \in A_1 \cup \ldots \cup A_n\} \cup \{W_{acc(a)} \mid a \in A_1 \cup \ldots \cup A_n\} \cup \{B_a \mid a \in A_1 \cup \ldots \cup A_n\}$ *and* $\longmapsto \subseteq MA \times MA$ *is a binary relation on MA such that hold the conditions of Definitions 9, 10, and 12, and:*

- $B_a \longmapsto X_{B_a,a}$ *iff* $a \in A_i$*, and* $X_{B_a,a} \longmapsto Y_{B_a,a}$ *iff* $a \in A_i$*, and* $Y_{B_a,a} \longmapsto W_{B_a,a}$ *iff* $a \in A_i$*, and* $W_{B_a,a} \longmapsto acc(a)$ *iff* $a \in A_i$ *and*
- $B_{a\rightarrow b} \longmapsto X_{B_{a\rightarrow b},a\rightarrow b}$ *iff* $a \rightarrow b \in \rightarrow_i$*, and* $X_{B_{a\rightarrow b},a\rightarrow b} \longmapsto Y_{B_{a\rightarrow b},a\rightarrow b}$ *iff* $a \rightarrow b \in \rightarrow_i$*, and* $Y_{B_{a\rightarrow b},a\rightarrow b} \longmapsto W_{B_{a\rightarrow b},a\rightarrow b}$ *iff* $a \rightarrow b \in \rightarrow_i$*, and* $W_{B_{a\rightarrow b},a\rightarrow b} \longmapsto Y_{a,b}$ *iff* $a \rightarrow b \in \rightarrow_i$ *and*
- $trust(i) \longmapsto X_{trust(i),W_{acc(a)}}$ *iff* $a \in A_i$*, and* $X_{trust(i),W_{acc(a)}} \longmapsto Y_{trust(i),W_{acc(a)}}$ *iff* $a \in A_i$*, and* $Y_{trust(i),W_{acc(a)}} \longmapsto W_{acc(a)}$ *iff* $a \in A_i$*, and* $W_{acc(a)} \longmapsto B_a$ *iff* $a \in A_i$*, and*
- $trust(i) \longmapsto X_{trust(i),W_{Y_{a,b}}}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$*, and* $X_{trust(i),W_{Y_{a,b}}} \longmapsto Y_{trust(i),W_{Y_{a,b}}}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$*, and* $Y_{trust(i),W_{Y_{a,b}}} \longmapsto W_{Y_{a,b}}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$*, and* $W_{Y_{a,b}} \longmapsto B_{a\rightarrow b}$ *iff* $a,b \in A_i$ *and* $a \rightarrow_i b$*, and*
- $acc(a) \longmapsto B_b$ *iff* $a \in A_i, b \in A_j$ *and* $aDT_{is}b$*, and*
- $acc(a) \longmapsto Y_{B_a,a}$ *iff* $a \in A_i, b \in A_j$ *and* $aDT_{ic}b$*, and*
- $acc(a) \longmapsto B_{b\rightarrow c}$ *iff* $a \in A_i, b,c \in A_j$ *and* $aDT_{is}(b \rightarrow_j c)$*, and*
- $acc(a) \longmapsto Y_{B_{b\rightarrow c},b\rightarrow c}$ *iff* $a \in A_i, b,c \in A_j$ *and* $aDT_{ic}(b \rightarrow_j c)$.

We say that an argument $a \in A_i$ *or attack* $a \rightarrow b \in \rightarrow_i$ *is untrustworthy concerning sincerity when there is an attack from an argument* $c \in A_j$ *to* B_a *or* $B_{a\rightarrow b}$*,* $cDT_{js}a$ *or* $cDT_{js}(a \rightarrow b)$*. We say that an argument* $a \in A_i$ *or attack* $a \rightarrow b \in \rightarrow_i$ *is untrustworthy concerning competence when there is an attack from an argument* $c \in A_j$ *to* $Y_{B_a,a}$ *or* $Y_{B_{a\rightarrow b},a\rightarrow b}$*,* $cDT_{jc}a$ *or* $cDT_{jc}(a \rightarrow b)$.

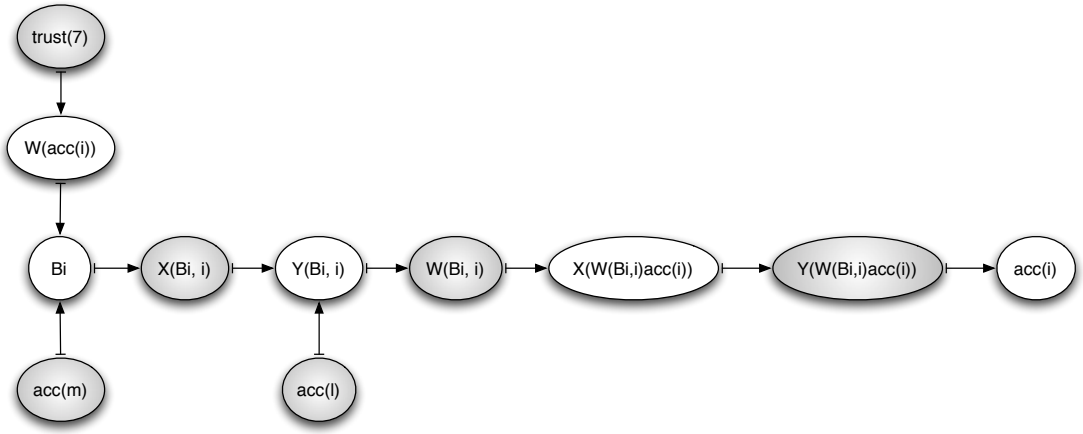

Figure 10: The flattening of the competence and sincerity's framework.

The flattening of the new framework distinguishing between attacks towards the sincerity of a source in proposing an information item, and attacks towards the competence of a source in proposing an information item is formalized in Definition 14. An example of flattening is visualized in Figure 10.

Example 5. *The meta-argument representing Witness7, $trust(7)$, supports by means of the auxiliary argument $W_{acc(i)}$ meta-argument B_i representing the fact that argument i is believed by Witness7. If this meta-argument is accepted, it means that there are no doubts about the sincerity of Witness7 concerning argument i. Meta-argument B_i supports meta-argument $acc(i)$, representing argument i in the meta-level. This support relation is built in the same way as the support between the sources and their information items, which means that meta-argument B_i attacks, towards auxiliary arguments X and Y, the auxiliary argument $W_{B_i,i}$. This auxiliary argument attacks, always by means of X and Y auxiliary arguments, the meta-argument $acc(i)$. In this framework, the acceptability of meta-argument $acc(i)$ depends on the acceptability of the belief regarding argument i. An attack towards the competence of argument i, instead, is addressed against meta-argument $Y_{B_i,i}$. In this way, argument $acc(i)$ can be made unacceptable in two ways: (1) by attacking directly meta-argument B_i (sincerity), and (2) by attacking the attack from B_i to $W_{B_i,i}$ (competence). Figure 10 shows these two cases with the attacks from argument m and argument l, respectively.*

Note that we do

Proposition 8. *Suppose a trustworthy source i provides evidence for argument*

a, and another trustworthy source j provides evidence for an argument b where b attacks the trustworthiness of argument a, then the extensions are the same if argument b attacks the sincerity or the competence of argument a.

Proof. Assume argument a and argument b are not attacked by other arguments. We use case analysis.

Case 1 : Let argument b attack the sincerity dimension of the trustworthiness of argument a. Meta-argument $acc(b)$ is accepted, as supported by a trustworthy source and not attacked by external arguments, and meta-argument $Y_{acc(b),B_a}$ is accepted. This means that meta-argument B_a is unaccepted, and thus argument $W_{B_a,a}$ is accepted, and meta-argument $acc(a)$ is unaccepted. Argument a is not part of any admissibility-based extension, and argument b is part of the admissibility-based extensions.

Case 2 : Let argument b attack the competence dimension of the trustworthiness of argument a. Meta-argument $acc(b)$ is accepted, as supported by a trustworthy source and not attacked by external arguments, and argument B_a is accepted, as supported by a trustworthy source. Then meta-argument $Y_{B_a,a}$ is unaccepted, as attacked by argument $acc(b)$, due to the conflict-free principle. This means that meta-argument $W_{B_a,a}$ is accepted, and meta-argument $acc(a)$ is unaccepted. Argument a is not part of any admissibility-based extension, and argument b is part of the admissibility-based extensions.

The extensions are the same whether argument b attacks the sincerity or the competence of argument a. □

5. Conclusions

Trust plays an important role in many research areas of artificial intelligence, particularly in the semantic web and multiagent systems where the sources have to deal with conflicting information from other sources. Building on the socio-cognitive model of trust described in [1], and on previous work integrating trust and argumentation [8], in this paper we presented a formal framework for modeling how different dimensions of the perceived trustworthiness of the source interact to determine the acceptability of the message, and how deviations from such expectation produce a specific feedback on source trustworthiness. Here, we applied this model to the case of sources exchanging and assessing arguments, but it could easily be extended to the exchange of any kind of factual information. The

reason why we focused first on argumentation is because this provides a window on the agent's reasoning.

In our model, the information sources are introduced into the argumentation framework. In argumentation systems as ASPIC+, arguments come from a single knowledge base and they have the form $\langle\{p, p \rightarrow q\}, q\rangle$. We propose to introduce the sources, e.g., $\langle\{1 : p, 2 : p \rightarrow q\}, 2 : q\rangle$, by instantiating abstract argumentation with the different knowledge bases of the sources using meta-argumentation. The sources have the form of meta-arguments $trust(s)$ in the meta-level. In this kind of representation, all the sources are considered as trustworthy as soon as their meta-argument are accepted in the meta-argumentation framework. A source which is considered untrustworthy is attacked by means of those argument(s) explicating that.

The arguments in the meta-argumentation framework need to be supported by some evidence in order to be accepted. This kind of evidence is represented in the model by means of an attack from the sources to the "default" meta-arguments $W(acc(arg_1))$ invalidating the acceptability of the arguments, e.g., arg_1. The evidence does not guarantee the acceptability of the arguments which can always be attacked by other arguments, but an argument may be accepted if and only if there is some evidence in favor of it.

The attacks to the trustworthiness of the sources can be focused on single arguments or attacks. In particular, when an argument attacks the trustworthiness of an information item, it is intended as an attack towards the item *as proposed by the source s*. We model this kind of attacks as an attack in the meta-level from the argument expressed against the trustworthiness of the information item to the attack of the source proposing the item towards the "default" meta-argument $W(acc(arg_1))$.

We introduce the feedback from the information sources to their items and converse. In such a way, when a source proposes a number of untrustworthy items which overcomes a given threshold, then also the source becomes untrustworthy. This is modelled as an attack from all the arguments proposed by a source s to a specific pattern which is "activated" after n attacks. The activation of the pattern leads to the activation of an attack from the pattern to the meta-argument $trust(s)$.

Finally, we distinguish two dimensions of trust, namely sincerity and competence. This separation allows us to distinguish also the attacks towards the trustworthiness of the information items. In particular, the sources provide an evidence in favor of the arguments by attacking indirectly meta-argument $acc(arg_1)$, with an attack against meta-argument B_{arg_1}. This meta-argument represents the beliefs of source s. Thus an attack against B_{arg_1} means that s is considered insincere con-

cerning argument arg_1. An attack against the support of B_{arg_1} to meta-argument $acc(arg_1)$ means that s is considered incompetent concerning argument arg_1.

We address several issues as future research.

Following Castelfranchi and Falcone [1], only a cognitive agent can "trust" another agent: only an agent endowed with goals and beliefs. In this model, first, one trusts another only relatively to a goal, i.e., for something s/he want to achieve, that s/he desires. Second, trust itself consists of beliefs. The first line of future work consists in representing also the actions and the goals in our model to follow this model. At time being, the implicit goal of the sources is to get their arguments accepted.

The arguments in this paper are treated basically as black boxes, as it is most often the case in works based on abstract argumentation, in the vein of [22]. This is significant in two respects. First, we did not discuss the two-way relationship between source trustworthiness and trust in the message when what is being communicated is not the argument as a whole, but rather one of its constituents, e.g., a premise, its conclusion, or the inference rule licensing the argument, as in [7]. Finding out that the source is mistaken on the truth of some premise (hence the argument is unsound) rather than on the truth of the inference (hence the argument is invalid) is likely to have very different effects for the feedback on the source, which will have to be investigated in future work. Second, we treat only the case of valid arguments, again as it is customary in abstract argumentation after [22]. This is of course a huge idealization with respect to everyday argumentation: as underlined by Walton [31], we rarely exchange deductively valid arguments, while the vast majority of arguments are defeasible, which implies a different sort of consequence relation.

Finally, the framework does not capture the cumulative effect of converging sources on argument acceptability. When more than one source offers the same information item, its acceptability is positively affected, as discussed in [32].

References

[1] C. Castelfranchi, R. Falcone, Trust Theory: A Socio-Cognitive and Computational Model, Wiley, 2010.

[2] J. Dix, S. Parsons, H. Prakken, G. R. Simari, Research challenges for argumentation, Computer Science - R&D 23 (1) (2009) 27–34.

[3] I. Rahwan, G. Simari (Eds.), Argumentation in Artificial Intelligence, Springer, 2009.

[4] H. Prade, A qualitative bipolar argumentative view of trust, in: SUM, 2007, pp. 268–276.

[5] R. Stranders, M. de Weerdt, C. Witteveen, Fuzzy argumentation for trust, in: CLIMA, 2007, pp. 214–230.

[6] P.-A. Matt, M. Morge, F. Toni, Combining statistics and arguments to compute trust, in: AAMAS, 2010, pp. 209–216.

[7] S. Parsons, Y. Tang, E. Sklar, P. McBurney, K. Cai, Argumentation-based reasoning in agents with varying degrees of trust, in: AAMAS, 2011, pp. 879–886.

[8] S. Villata, G. Boella, D. M. Gabbay, L. van der Torre, Arguing about the trustworthiness of the information sources, in: W. Liu (Ed.), ECSQARU, Vol. 6717 of Lecture Notes in Computer Science, Springer, 2011, pp. 74–85.

[9] C. da Costa Pereira, A. Tettamanzi, S. Villata, Changing ones mind: Erase or rewind?, in: T. Walsh (Ed.), IJCAI, IJCAI/AAAI, 2011, pp. 164–171.

[10] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artif. Intell. 77 (2) (1995) 321–357.

[11] H. Prakken, An abstract framework for argumentation with structured arguments, Tech. Rep. UU-CS-2009-019, Utrecht University (2009).

[12] T. F. Gordon, H. Prakken, D. Walton, The Carneades model of argument and burden of proof, Artif. Intell. 171 (10-15) (2007) 875–896.

[13] H. Jakobovits, D. Vermeir, Robust semantics for argumentation frameworks, J. Log. Comput. 9 (2) (1999) 215–261.

[14] S. Modgil, T. Bench-Capon, Metalevel argumentation, Tech. rep., www.csc.liv.ac.uk/research/ techreports/techreports.html (2009).

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

[16] C. Cayrol, M.-C. Lagasquie-Schiex, Coalitions of arguments: A tool for handling bipolar argumentation frameworks, Int. J. Intell. Syst. 25 (1) (2010) 83–109.

[17] S. Villata, Meta-argumentation for multiagent systems: Coalition formation, merging views, subsumption relation and dependence networks, Ph.D. thesis, University of Turin (2010).

[18] P. Baroni, M. Giacomin, On principle-based evaluation of extension-based argumentation semantics, Artif. Intell. 171 (10-15) (2007) 675–700.

[19] G. Boella, L. van der Torre, S. Villata, On the acceptability of meta-arguments, in: Procs. of IAT, IEEE, 2009, pp. 259–262.

[20] S. Coste-Marquis, C. Devred, S. Konieczny, M.-C. Lagasquie-Schiex, P. Marquis, On the merging of Dung's argumentation systems, Artif. Intell. 171 (10-15) (2007) 730–753.

[21] S. Parsons, P. McBurney, E. Sklar, Reasoning about trust using argumentation: A position paper, in: Procs. of ArgMAS, 2010.

[22] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artif. Intell. 77 (2) (1995) 321–358.

[23] Y. Tang, K. Cai, E. Sklar, P. McBurney, S. Parsons, A system of argumentation for reasoning about trust, in: EUMAS, 2010.

[24] D. Gambetta, Can we trust them?, Trust: Making and breaking cooperative relations (1990) 213–238.

[25] E. Lorini, R. Demolombe, From binary trust to graded trust in information sources: A logical perspective, in: R. Falcone, K. S. Barber, J. Sabater-Mir, M. P. Singh (Eds.), AAMAS-TRUST, Vol. 5396 of Lecture Notes in Computer Science, Springer, 2008, pp. 205–225.

[26] C.-J. Liau, Belief, information acquisition, and trust in multi-agent systems–a modal logic formulation, Artif. Intell. 149 (1) (2003) 31–60.

[27] Y. Wang, M. P. Singh, Formal trust model for multiagent systems, in: M. M. Veloso (Ed.), IJCAI, 2007, pp. 1551–1556.

[28] A. Bochman, A causal approach to nonmonotonic reasoning, Artif. Intell. 160 (1-2) (2004) 105–143.

[29] G. Boella, D. M. Gabbay, L. van der Torre, S. Villata, Support in abstract argumentation, in: Procs. of COMMA, IOS Press, 2010, pp. 40–51.

[30] S. Villata, G. Boella, L. van der Torre, Argumentation patterns, in: Proceedings of the 8th International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2011), 2011, pp. 133–150.

[31] D. Walton, C. Reed, F. Macagno, Argumentation Schemes, CUP, 2008.

[32] C. Castelfranchi, Representation and integration of multiple knowledge sources: issues and questions, in: V. Cantoni, V. Di Gesù, A. Setti, D. Tegolo (Eds.), Human & Machine Perception: Information Fusion, Plenum Press, 1997.