A Dynamic Model of Trust in Dialogues

Gideon Ogunniye[1(✉)], Alice Toniolo[2], and Nir Oren[1]

[1] Department of Computing Science,
University of Aberdeen, Aberdeen, Scotland, UK
g.ogunniye@abdn.ac.uk
[2] School of Computer Science, University of St Andrews,
St. Andrews, Scotland, UK

Abstract. In human interactions, trust is regularly updated during a discussion. For example, if someone is caught lying, any further utterances they make will be discounted, until trust is regained. This paper seeks to model such behaviour by introducing a dialogue game which operates over several iterations, with trust updates occurring at the end of each iteration. In turn, trust changes are computed based on intuitive properties, captured through three rules. By representing agent knowledge within a preference-based argumentation framework, we demonstrate how trust can change over the course of a dialogue.

1 Introduction

Within a dialogue, participants exchange arguments, aiming to achieve some overarching goals. Typically, these participants have partial information and individual preferences and goals, and the parties aim to achieve an outcome based on these individual contexts. Importantly, some dialogue participants may be malicious or incompetent, and—to achieve desirable dialogical outcomes—the inputs from these parties should be discounted. In human dialogues, such participants are characterised by the lack of trust ascribed to them, and in this work we consider how such trust should be computed.

While previous work [12] has considered how the trust of participants should be updated *following* a dialogue, we observe that in long-lasting human discussions, trust can change during the dialogue itself. For example, within a courtroom, a witness who repeatedly appears to lie will not be believed even if they later act honestly. Trust can be viewed as making the arguments of more trusted agents be preferred—in the eyes of those observing the dialogue—to the arguments of less trusted agents. Importantly, there appears to be a feedback cycle at play within dialogue: low trust in a dialogue participant can lead to further reductions of trust as they are unable to provide sufficient evidence to be believed. To accurately model dialogue and reason about the trust ascribed to its participants, it is critical to take this feedback cycle between utterances and trust into account. This paper considers such a feedback cycle.

E. Black et al. (Eds.): TAFA 2017, LNAI 10757, pp. 211–226, 2018.
https://doi.org/10.1007/978-3-319-75553-3_15

The research questions we address in this work are as follows. (1) How should trust change during the course of a dialogue based on the utterances made by dialogue participants? (2) How should trust affect the justified conclusions obtained from a dialogue?

To answer these questions, we describe a dialogue model in which participants interact by exchanging arguments. Within this model, we define a trust relation for each participant with respect to other participants (encoded as a preference ordering over the participants), and describe how each participant updates its trust relation. In particular, each participant observes the behaviours of others and uses these observations as an input to update its trust relation (for the other participants) through a trust update function.

To compute the justified conclusions of a dialogue, we instantiate a preference-based argumentation framework (PAF) [1]. As a result, each participant can identify its own set of preferred conclusions, and a set of justified conclusions can be identified from these sets.

The proposed framework permits us to better represent the feedback relationship between trust and dialogue. The remainder of the paper is organised as follows: Sect. 2 recalls preference-based argumentation frameworks [1] and provides a brief overview of our notion of trust in dialogues. Section 3 describes our proposed dialogue model. Section 4 describes the trust update rules and the process we considered for dynamically updating trust within our dialogue model. Section 5 describes how the preference-based argumentation framework is instantiated in our model. Section 6 illustrates how trust update rules are applied through an example. Section 7 compares our approach with some existing works. Section 8 presents our conclusions and some directions for future work.

2 Background

Preference-based argumentation frameworks extend abstract argumentation frameworks [7], and we therefore begin by describing the former.

Definition 1. *An Argumentation Framework $\mathcal{F}$ is defined as a pair $\langle \mathcal{A}, \mathcal{R} \rangle$ where $\mathcal{A}$ is a set of* arguments *and $\mathcal{R}$ is a binary attack relation on $\mathcal{A}$.*

Extensions are sets of arguments that are, in some sense, justified. These extensions are computed using one of several *argumentation semantics*.

Preference-based argumentation frameworks [1] seek to capture the relative strengths of arguments and can be instantiated in different ways. In this paper, we will use preference-based argumentation frameworks to encode trust in other dialogue participants, allowing us to compute which arguments should, or should not be considered justified.

Within a preference-based argumentation framework, preferences are encoded through a reflexive and transitive binary relation $\geq$ over the arguments of $\mathcal{A}$. Given two arguments $\phi_1, \phi_2 \in \mathcal{A}$, $\phi_1 \geq \phi_2$ means that ϕ_1 is at least as preferred as ϕ_2. The relation $>$ is the strict version of $\geq$ i.e., $\phi_1 > \phi_2$ iff $\phi_1 \geq \phi_2$ but $\phi_2 \not\geq \phi_1$. As usual, $\phi_1 = \phi_2$ iff $\phi_1 \geq \phi_2$ and $\phi_2 \geq \phi_1$.

Given this, a preference-based argumentation framework is defined as follows.

Definition 2. *A Preference-based argumentation framework (PAF for short) [1] is a tuple* $\mathcal{T} = \langle \mathcal{A}, \mathcal{R}, \geq \rangle$ *where* $\mathcal{A}$ *is a set of arguments,* $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ *is an attack relation and* $\geq \subseteq \mathcal{A} \times \mathcal{A}$ *is a (partial or total) preorder on* $\mathcal{A}$*. The extensions of* $\mathcal{T}$ *under a given semantics are the extensions of the argumentation framework* $(\mathcal{A}, \mathcal{R}_r)$*, called the* repaired framework*, under the same semantics with* $\mathcal{R}_r = \{(\phi_1, \phi_2) | (\phi_1, \phi_2) \in \mathcal{R} \text{ and } (\phi_2 \not> \phi_1)\} \bigcup \{(\phi_2, \phi_1) | (\phi_1, \phi_2) \in \mathcal{R} \text{ and } \phi_2 > \phi_1\}$.

Given a PAF, one can identify different sets of justified conclusions by considering different extensions. PAFs extend standard Dung argumentation frameworks with the addition of preferences between arguments to repair *critical attacks* and refine the extension of the repaired PAF. Therefore, we also define the semantics of standard argumentation frameworks, the notion of critical attacks and extension refinement. In this paper we will focus on the preferred semantics.

Definition 3. *Given* $\mathcal{F} = \langle \mathcal{A}, \mathcal{R} \rangle$*, a set of arguments* $\mathcal{E} \subseteq \mathcal{A}$ *is said to be* conflict-free *iff* $\forall \phi_1, \phi_2 \in \mathcal{E}$*, there is no* $(\phi_1, \phi_2) \in \mathcal{R}$*. Given an argument* $\phi_1 \in \mathcal{E}$*,* $\mathcal{E}$ *is said to* defend ϕ_1 *iff for all* $\phi_2 \in \mathcal{A}$*, if* $(\phi_2, \phi_1) \in \mathcal{R}$ *then there is a* $\phi_3 \in \mathcal{E}$ *such that* $(\phi_3, \phi_2) \in \mathcal{R}$*.* $\mathcal{E}$ *is admissible iff it is conflict-free and defends all its elements.* $\mathcal{E}$ *is a complete extension iff there are no other arguments which it defends.* $\mathcal{E}$ *is a preferred extension iff it is a maximal (with respect to set inclusion) complete extension.*

Preferred semantics admit multiple extensions; here, such an extension represents a potentially justified view (which conflicts with other views). If an argument is present in all extensions, then it is *sceptically* justified; while if it is present in at least one extension, it is *credulously* justified.

Definition 4. *(Critical attack) [1]. Let* $\mathcal{F}$ *be an argumentation framework and* $\geq \subseteq \mathcal{A} \times \mathcal{A}$*. An attack* $(\phi_2, \phi_1) \in \mathcal{R}$ *is* critical *iff* $\phi_1 > \phi_2$*.*

PAFs repair critical attacks on the graph of attacks by *inverting* the arrow of the attack relation (i.e., $(\phi_2, \phi_1) \in \mathcal{R}$ with $\phi_1 > \phi_2$ becomes $(\phi_1, \phi_2) \in \mathcal{R}$). This repair property ensures that arguments that are more preferred in an argumentation framework *defeat* arguments that are less preferred. An argument ϕ_1 *defeats* ϕ_2 *iff* $((\phi_1, \phi_2)$ or $(\phi_2, \phi_1)) \in \mathcal{R}$ and $\phi_1 > \phi_2$. For a symmetric attack relation, removing critical attacks gives the same results as inverting attacks. Extensions are then constructed from the corresponding repaired PAF using the semantics of $\mathcal{F}$. In addition, in PAFs, a refinement relation is used to refine the results of a framework by comparing its extensions.

Definition 5. *(Refinement relation) [1]. Let* $(\mathcal{A}, \geq)$ *be such that* $\mathcal{A}$ *is a set of arguments and* $\geq \subseteq \mathcal{A} \times \mathcal{A}$ *is a (partial or total) preorder. A refinement relation denoted by* $\geq_r$*, is a binary relation on* $\mathcal{P}(\mathcal{A})^2$ *such that* $\geq_r$ *is reflexive, transitive and for all* $\mathcal{E} \subseteq \mathcal{A}$*, for all* $\phi_1, \phi_2 \in \mathcal{A} \backslash \mathcal{E}$*, if* $\phi_1 > \phi_2$ *then* $\mathcal{E} \bigcup \{\phi_1\} >_r \mathcal{E} \bigcup \{\phi_2\}$.

Let *Ags* be a set of participants within a dialogue. We consider that each dialogue participant $Ag_{i \in Ags}$, for $i = 1, \ldots, n$, has an associated trust relation over other participants, encoded through a preference ordering $\succeq_{Ag_i}$.

Definition 6. *Let Ags be a set of dialogue participants. The trust relation of a given participant Ag_i over Ags is a preference ordering $\succeq_{Ag_i} \subseteq Ags \times Ags$. $Ag_j \succeq_{Ag_i} Ag_k$ denotes that Ag_i prefers (trusts) Ag_j to Ag_k.*

We consider the following properties for the trust relation:

- *Non-Symmetric:* if a participant Ag_i trusts another participant Ag_j, this does not imply that Ag_j trusts Ag_i.
- *Transitive:* Unlike some other works on trust [9,17], we assume that transitivity of trust (also known as *derived trust*) is not required in our model. As a result, we assume that a given participant has the ability to decide whether or not to trust another participant at any stage of the dialogue.

The trust relation represents the viewpoint of a given participant independently of the trust relations of other participants. Therefore, unlike the systems described in, for example, [9,17], there is no need to represent a 'global map' of trust relations—a trust network—in our model.

3 A Formal Dialogue Model

We consider a dialogue system where each participant Ag_i has two main components: a *knowledge base* (containing its trust relation over other participants, a set of arguments, and a set of attacks between arguments) and a *commitment store*. We follow Hamblin (as cited in [19]) in defining a commitment store as a "store of statements" that represents the arguments a participant is publicly committed to.

Definition 7. *The knowledge base of a participant $Ag_i \in Ags$ is a tuple $\mathcal{KB}_{Ag_i} = \langle A_{Ag_i}, R_{Ag_i}, \succeq_{Ag_i} \rangle$, where A_{Ag_i} is the set of arguments known by Ag_i (representing their own knowledge); $R_{Ag_i} \subseteq A_{Ag_i} \times A_{Ag_j}$ is a set of attacks where $(\phi_1, \phi_2) \in R_{Ag_i}$ iff $\phi_1 \in A_{Ag_i}$ and ϕ_2 is an argument provided by any participant Ag_j; and $\succeq_{Ag_i}$ is the trust relation (c.f., Definition 6) of Ag_i with regards to other participants.*

Each participant updates its knowledge base at the end of each iteration of a dialogue. Intuitively, an iteration represents a subdialogue, including an exchange of arguments arising from a participant's (potentially) controversial assertion. Unlike the knowledge base, the commitment store is updated after every dialogue move made by the participant.

Definition 8. *The commitment store of a participant $Ag_i \in Ags$ at iteration $t \in \{1 \ldots n\}$ is a set $CS^t_{Ag_i} = \{\phi_1, \ldots, \phi_n\}$ which contains arguments introduced into the dialogue by Ag_i at iteration t such that $CS^0_{Ag_i} = \emptyset$.*

The union of the commitment stores of all participants is called the *universal commitment store* $\mathcal{UCS}^t = \bigcup_{Ag_i} CS^t_{Ag_i}$. An argument put forward by a participant may be attacked by an argument from another participant. Therefore, in our dialogue system, an argumentation framework $\langle \mathcal{UCS}^t, \mathcal{R} \rangle$ is induced by the set of arguments exchanged during dialogue in the universal commitment store and their respective attacking relationships as in [7]. Hence, $(\phi_1, \phi_2) \in \mathcal{R}$ if $(\phi_1, \phi_2) \in R_{Ag_i}$, $\phi_1 \in CS_{Ag_i}$ and $\phi_2 \in \mathcal{UCS}^t$. The universal commitment store can be viewed as the global state of the dialogue at a given iteration.

We now turn our attention to the dialogue game itself. A dialogue game like the one described in [13] specifies the major elements of a dialogue, such as its commencement, combination, and termination rules among others. Likewise, the system described in [11] specifies how the topic of discussion in a dialogue can be represented in some logical language. We are interested in how a participant updates its commitment store and its trust relation in a dialogue when it, or other participants, introduce arguments. We assume that at iteration t, a participant is allowed to add arguments to its commitment store if it is not already present within the store (and was not previously present), and retract arguments from its commitment store only if the argument was already present in the store.

3.1 Protocol Rules and Speech Acts

Protocol rules regulate the set of legal moves that are permitted at each iteration of a dialogue. In our framework, a dialogue consists of multiple discrete iterations t within which the moves are made. A dialogue move is referred to as M^t_x where $x, t \in \mathbb{N}$, denoting that a move with identifier x is made at iteration t. At its most general, a protocol identifies a legal move based on all previous dialogue moves.

Definition 9. *A* dialogue D *consists of a sequence of* iterations *such that* $D = [[M^1_1, \ldots, M^1_x], \ldots, [M^t_1, \ldots, M^t_x]]$. *The dialogue involves* n *participants* $Ag_1, \ldots, Ag_n$ *where* $(n \geq 2)$. *Within a dialogue* D, *iteration* j *consists of a sequence of moves* $[M^j_1, \ldots M^j_x]$.

A dialogue participant evaluates the set of arguments exchanged within an iteration to update its trust relation over other participants. Within each iteration, there is a claim to be discussed and arguments that attack or defend the claim. Note that a claim is abstractly represented as an argument. An iteration therefore represents a sub-discussion focused around a single topic of the overarching dialogue, which can be treated in an atomic manner with regards to trust.

The dialogue protocol is as described in Fig. 1. Each node—except the 'update' node (described in detail later)—represents a speech act, and the outgoing arcs from a node indicate possible responding speech acts. We consider four types of speech acts, denoted $assert(Ag_i, \phi, t)$, $contradict(Ag_i, \phi_1, \phi_2, t)$, $retract(Ag_i, \phi, t)$, and *exit* respectively. A participant Ag_i uses $assert(Ag_i, \phi, t)$ to put forward a claim $\phi \in A_{Ag_i}$ at iteration t. A $contradict(Ag_i, \phi_1, \phi_2, t)$ move attacks a previous argument $\phi_1 \in A_{Ag_j}$ from another participant Ag_j by argument $\phi_2 \in A_{Ag_i}$ from participant Ag_i. A participant Ag_i uses $retract(Ag_i, \phi, t)$ to retract its previous argument. A participant uses *exit* to exit an iteration. This move is made

when a participant has no more arguments to advance within the iteration. When an iteration concludes (shown by the terminal *update* node in the figure), trust is updated. The dialogue then proceeds to the next iteration, or may terminate. A dialogue therefore consists of at least one, but potentially many more, iterations.

In addition to the constraints on the type of speech act that can be made in a dialogue, we also consider the *relevance* of a move. A move M^t_{x+i}, for $x, i \geq 1$ is *relevant* to iteration t if the argument of the move will affect the justification of the argument of the move M^t_x. Specifically, an argument ϕ_2 in move M^t_{x+i} affects the justification of an argument ϕ_1 in M^t_x if it attacks ϕ_1 (c.f., [14]). Relevance is defined from the second move of an iteration (i.e., when $x \geq 1$) because the first move is taken to introduce the claim to be discussed in the iteration. The protocol rules enforce that ϕ_2 is relevant to an iteration t if it affects the justification of ϕ_1 that has been previously moved in the iteration. However, if ϕ_1 is retracted in the iteration, ϕ_2 is no longer relevant and must be retracted except if it affects the justification of another argument ϕ_3. Furthermore, as the outgoing arcs in Fig. 1 depict, a move to exit an iteration is also considered relevant from the second move but a move to retract an argument is only considered relevant from the third move (i.e., when $x \geq 2$). These constraints help to prevent participants from making moves that are not relevant to the current iteration.

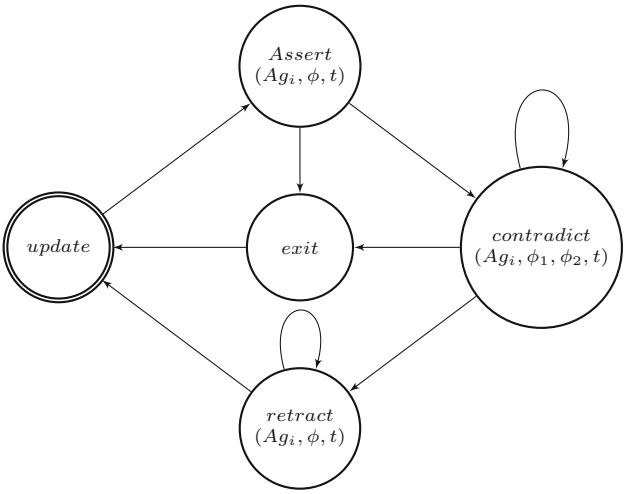

Fig. 1. Protocol rules

3.2 Commitment Rules

A participant's commitment store is revised throughout the dialogue as it advances arguments. Therefore, it is important to define how each of the proposed speech acts updates a participant's commitment store.

Definition 10. *The commitment store of a participant $Ag_i \in Ags$ is updated as follows:*

$$CS^t_{Ag_i} = \begin{cases} \emptyset & \textit{iff } t = 0, \\ CS^{t-1}_{Ag_i} \bigcup \{\phi\} & \textit{iff } m^t_x = assert(Ag_i, \phi, t), \\ CS^{t-1}_{Ag_i} \bigcup \{\phi_2\} & \textit{iff } m^t_x = contradict(Ag_i, \phi_1, \phi_2, t) \\ CS^{t-1}_{Ag_i} \setminus \{\phi\} & \textit{iff } m^t_x = retract(Ag_i, \phi, t) \\ CS^{t-1}_{Ag_i} & \textit{iff } m^t_x = exit \end{cases}$$

4 Updating Trust

We now turn our attention to how trust should be updated as a dialogue progresses. We limit our focus to how the trust relation component of a participant's knowledge base ($\succeq_{Ag_i}$) is updated. A trust update function is used to perform this update when an iteration concludes, as represented by the 'update' node in Fig. 1.

As input, the trust update function takes a participant's *trust update rules* and its *preference on the trust update rules*. In the remainder of this section, we formalise both of these concepts.

Trust update rules describe the situations in which trust in a dialogue participant should change. In this paper, we consider the following trust update rules.

- A dialogue participant whose arguments are self-contradicting should be less trusted than a consistent participant.
- A dialogue participant who is unable to justify its arguments should be less trusted than one who can.
- A dialogue participant who regularly retracts arguments should be less trusted than one who does not.

These rules are similar to some of the properties that have been considered in the literature of ranking-based semantics for abstract argumentation (for a review on ranking-based semantics for abstract argumentation, see [3]). These rules are also supported by extension-based semantics (i.e., Dung's semantics [7]). For instance, the second rule could be represented as a participant having an argument ϕ in its commitment store, but not within an extension: $\phi \notin \mathcal{E}(\langle \mathcal{UCS}, \mathcal{R} \rangle)$[1]. We do not claim that the three trust update rules considered in this paper are exhaustive, and intend to investigate additional rules, taken from sources such as [3], in the future. We formalise the three trust update rules as follows.

Definition 11. *Self Contradicting Arguments (SC): A participant Ag_i is self contradicting if CS_{Ag_i} is not conflict free.*

[1] Here, $\mathcal{E}$ represents the extension(s) obtained on the argumentation framework $\langle \mathcal{UCS}, \mathcal{R} \rangle$.

Definition 12. *Lack of Justification (LJ): A participant Ag_i lacks justification for an argument ϕ_1 iff $\phi_1 \in CS_{Ag_i}$ and there is a $\phi_2 \in \mathcal{UCS} \backslash CS_{Ag_i}$ such that ϕ_2 defeats ϕ_1.*

Defeats consider preferences among attacks and are defined in Sect. 2.

Definition 13. *Argument Retraction (AR): A participant Ag_i is inconsistent iff $\phi_1 \in CS_{Ag_i}$ and there is a $\phi_2 \in \mathcal{UCS} \backslash CS_{Ag_i}$ such that ϕ_2 attacks ϕ_1 and Ag_i retracts ϕ_1 from CS_{Ag_i}.*

This rule also requires that if ϕ_2 attacks ϕ_1 and ϕ_1 is retracted by Ag_i, Ag_j is expected to retract ϕ_2 as enforced by the dialogue protocol without any loss of trust for Ag_j except if ϕ_2 attacks another argument ϕ_3 that is not retracted.

Given the three trust update rules considered, there are four possible combinations of these rules in an iteration. These possible combinations are given below.

- *(SC, LJ, AR)*: This combination means all the three trust updates rules occur within a particular iteration under consideration.
- *(SC, LJ)*: This combination means *self contradiction* and *lack of justification* occur within a particular iteration under consideration.
- *(SC, AR)*: This combination means *self contradiction* and *argument retraction* occur within a particular iteration under consideration.
- *(LJ, AR)*: This combination means *lack of justification* and *argument retraction* occur within a particular iteration under consideration.

Note that within an iteration, the arrangement of trust update rules in a combination is not important. For instance, *(SC, AR)* and *(AR, SC)* is considered to be the same combination.

Agents have preferences over trust update rules. For example, one may trust somebody who contradicts themselves much less than they trust someone who regularly retracts arguments. Such preferences on trust update rules are a partial order over trust update rules. This partial order specifies the order of importance a given participant attaches to the trust update rules.

Definition 14. *Let $TR^t_{Ags} = \{SC, LJ, AR\}$ be a set of trust update rules for the set of participants Ags at iteration t. A given participant's preference on TR^t_{Ags} is a partial ordering $\succeq^t_{Ag_{i(TR)}}$ such that for rules $X, Y \in TR^t_{Ags}$, $X \succeq^t_{Ag_i(TR)} Y$ denotes rule X has preference over rule Y in $\succeq^t_{Ag_{i(TR)}}$.*

Since we are concerned with the viewpoint of a given participant, dialogue participants may have varying preferences on trust update rules. Furthermore, such preferences may change from one iteration to another. For instance, in a particular iteration, a given participant may consider argument retraction as the least inconsistent behaviour if a target participant retracts an argument from its commitment store as a result of learning from the arguments of other participants that the retracted argument is inaccurate. This may not be the case

if the target participant is forced to retract an argument from its commitment store as a result of its inability to advance other arguments to defend it.

If the preference on the trust update rules of a given participant Ag_i is $\succeq_{Ag_{i(TR)}} = (SC \succ_{Ag_{i(TR)}} LJ \succ_{Ag_{i(TR)}} AR)$, then, *self contradiction* is most important when updating the participant's trust relation, followed by *lack of justification* and *argument retraction* respectively.

Consider a dialogue participant Ag_i, with a trust update function denoted by $\mathcal{UF}$ at *iteration* t of a dialogue. The participant exchanges arguments with other participants in the dialogue through defined *speech acts* and *protocol rules*. It updates its *commitment store* $CS^t_{Ag_i}$ after each of its *moves* m^t_x in the dialogue. It observes some *trust updates rules* based on the observed behaviours of other participants in a particular iteration of the dialogue. As earlier stated, the commitment store of all dialogue participants is publicly observable. Ag_i updates its *trust relation* $\succeq_{Ag_i}$ over other participants based on its *trust update rules* and *preference on the rules* $\succeq^t_{Ag_{i(TR)}}$, repeating the process in the next iteration.

We formalise the trust update function as follows.

Definition 15. *Let* $TR^t_{Ag_i}$ *be the trust update rules of a given participant* Ag_i*;* $\succeq^t_{Ag_{i(TR)}}$ *be the participant's preference on the trust update rules; and* $\succeq^t_{Ag_i}$ *its trust relation over other participants at iteration* $t \in \{1 \ldots n\}$*. The trust update function* $\mathcal{UF}$ *is a function of the form* $\mathcal{UF}$*:* $(TR^t_{Ag_i} \times \succeq^t_{Ag_{i(TR)}}) \rightarrow \succeq^t_{Ag_i}$ *which takes in* Ag_i*'s trust update rules and current trust preferences, and returns an updated set of trust preferences.*

A given participant's trust relation over other participants is updated via the trust update function. Such a relation provides the basis for computing what the participant deems justified in an iteration.

In the next section, we analyse how each participant computes extensions in their personalised preference-based argumentation frameworks.

5 Dialogue Outcome

Given an argumentation framework induced by the set of arguments exchanged during dialogue in the universal commitment store and their respective attacking relationships. Also, given a preference ordering over dialogue participants, we instantiate a PAF by providing a rational basis for the preferences between arguments. We prefer arguments $\phi_1 \geq \phi_2$ (or strictly prefer arguments $\phi_1 > \phi_2$) iff there are some dialogue participants Ag_i and Ag_j such that $\phi_1 \in CS_{Ag_i}, \phi_2 \in CS_{Ag_j}$ and $Ag_i \succeq Ag_j$ (respectively $Ag_i \succ Ag_j$). If there are critical attacks in $\langle \mathcal{UCS}, \mathcal{R} \rangle$, the attacks are repaired (c.f., Sect. 2). Moreover, the extensions generated from the $\langle \mathcal{UCS}, \mathcal{R} \rangle$ are refined as shown in Sect. 2.

Since the preference orderings over dialogue participants represent the viewpoint of a given participant in our model, it is possible to have as many preference orderings over participants as the number of participants in a dialogue. By implication, the notions of preferences between arguments; critical attacks; and

argument defeat are relative to each participant. In what follows, we introduce the notion of a *participant* for a *PAF* similar to the notion of an *audience* in [2]. Participants are individuated by their preferences over other dialogue participants leading to their preferences between arguments. The arguments in the $\mathcal{UCS}$ will then be evaluated by each participant in accordance with its preferences between arguments. This leads to the following argument framework.

Definition 16. *Let Ags be a set of participants* $\{Ag_1, \ldots, Ag_n\}$ *then for* $i = 1, \ldots, n$, *the* preference-base argumentation framework *of participant* Ag_i *is a tuple* $\mathcal{T}_{Ag_i} = \langle \mathcal{A}, \mathcal{R}, \succeq^{\mathcal{A}}_{Ag_i} \rangle$ *where* $\mathcal{A} \subseteq \mathcal{UCS}$ *is a set of arguments,* $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ *is an attack relation and* $\succeq^{\mathcal{A}}_{Ag_i} \subseteq \mathcal{A} \times \mathcal{A}$ *is a (partial or total) preorder on* $\mathcal{A}$ *according to* Ag_i.

An attack succeeds in the preference-based argumentation framework of a participant if it is not a critical attack or if the participant has no preference between the arguments. Thus, the set of *defeat relations* (attacks that succeed) in one participant's context may be different from the one in another participant's context. An argument $\phi_1 \in \mathcal{A}$ *defeats* another argument $\phi_2 \in \mathcal{A}$ *iff* $(\phi_1, \phi_2) \in \mathcal{R}$ and $\phi_2 \not\succeq^{\mathcal{A}}_{Ag_i} \phi_1$. Further, note that the *preferred semantics* of $\mathcal{T}_{Ag_i}$ may return a different refined preferred extension $\mathcal{E}_{Ag_i}$ to the *preferred semantics* of $\mathcal{T}_{Ag_j}$.

Definition 17. *A set of arguments* $\mathcal{E}_{Ag_i}$ *in a preference-based argumentation framework* $\mathcal{T}_{Ag_i}$ *is a preferred extension for a participant* Ag_i *if it is maximal (with respect to set inclusion) complete extension obtained from* $\mathcal{T}_{Ag_i}$.

To define the set of justified conclusions in our model, we borrow the notions of *objectively acceptable* and *subjectively acceptable* arguments from [2].

Definition 18. *Given a preference-based argumentation framework* $\mathcal{T}_{Ags} = \langle \mathcal{A}, \mathcal{R}, \succeq^{\mathcal{A}}_{Ags} \rangle$ *for some participants Ags, an argument* ϕ *is objectively acceptable iff for all* $Ag_i \in Ags$, ϕ *is in every* $\mathcal{E}_{Ag_i}$. *On the other hand,* ϕ *is subjectively acceptable iff for some* $Ag_i \in Ags$, ϕ *is in some* $\mathcal{E}_{Ag_i}$.

In the discussion thus far, we have shown that each dialogue participant computes its preferred extensions in a dialogue based on preference ordering (i.e., trust) over the other dialogue participants—leading to preference ordering over arguments. It then follows that out of the set of preferred extensions a given participant may have, the refined preferred extension is the extension whose arguments are more trusted than the other extensions in the set. Consequently, the set of objectively acceptable arguments is the set that the participants simultaneously considered as the most trusted set of arguments in the dialogue. We consider this set as the most justified conclusion of a dialogue similar to how the set of sceptically justified arguments is considered as the set of most justified arguments in standard argumentation frameworks and PAF. With this property, we show how trust can have an effect on the justified conclusions of a dialogue.

Next, we consider the notion of a cycle within the preference ordering.

Definition 19. *A preference-based argumentation framework* $\mathcal{T}_{Ags} = \langle \mathcal{A}, \mathcal{R}, \succeq^{\mathcal{A}}_{Ags} \rangle$ *for participants Ags has a cycle iff there are two arguments* $\phi_1, \phi_2 \in \mathcal{A}$ *such that* $\phi_1 \succeq^{\mathcal{A}}_{Ags} \phi_2$ *and* $\phi_2 \succeq^{\mathcal{A}}_{Ags} \phi_1$.

Proposition 1. *Assume preferred semantics, for any* $\mathcal{T}_{Ag_i}$*, if* $(\phi_1, \phi_2) \in \mathcal{R}$ *and* $\phi_2 \succ^{\mathcal{A}}_{Ag_i} \phi_1$*, then* ϕ_1 *is not accepted—*$\phi_1 \notin \mathcal{E}_{Ag_i}$*.*

Proof. For any $\mathcal{T}_{Ag_i}$ that is cycle free, there is a unique corresponding $\mathcal{F}$, $\mathcal{F}_{Ag_i} = \langle \mathcal{A}, \mathcal{R} \rangle$, such that an element of attack relation $(\phi_1, \phi_2) \in \mathcal{R}$ in $\mathcal{F}_{Ag_i}$ is an element of defeat relation $(\phi_1, \phi_2) \in \mathcal{R}$ in $\mathcal{T}_{Ag_i}$. Therefore, the preferred extension of $\mathcal{F}_{Ag_i}$ will contain the same arguments as the preferred extension of $\mathcal{T}_{Ag_i}$. If $\mathcal{T}_{Ag_i}$ is cycle free, it means there is a preference ordering $\succeq^{\mathcal{A}}_{Ags}$ over $\mathcal{A}$. For $\phi_1, \phi_2 \in \mathcal{A}$, $(\phi_1, \phi_2) \in \mathcal{R}$ and $\phi_2 \succ^{\mathcal{A}}_{Ag_i} \phi_1$. The attack from ϕ_1 to ϕ_2 will be inverted. Therefore, this attack will not appear in $\mathcal{F}_{Ag_i}$. Instead, an attack from ϕ_2 to ϕ_1 will appear and since attack from ϕ_1 to ϕ_2 is not in $\mathcal{F}_{Ag_i}$, ϕ_2 is accepted in a preferred extension of $\mathcal{F}_{Ag_i}$ and ϕ_1 rejected. This applies to $\mathcal{T}_{Ag_i}$ since $\mathcal{T}_{Ag_i}$ corresponds to $\mathcal{F}_{Ag_i}$.

Proposition 2. *Suppose* $\mathcal{T}_{Ag_i}$ *has a cycle between all arguments (i.e.,* $(\forall \phi_1, \phi_2 \in \mathcal{A})$ *s.t.*$(\phi_1, \phi_2) \in \mathcal{R}$*,* $\phi_1 =^{\mathcal{A}}_{Ag_i} \phi_2$*), then any extension of* $\mathcal{T}_{Ag_i}$ *is also an extension of Dung's framework* $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ *and vice versa under the same semantics.*

Proof. This follows from Definition 2 and Proposition 1.

This property ensures that when Ag_i has equal or no preferences for some arguments in $\mathcal{T}_{Ag_i}$, then there can be no critical attacks between these arguments and preferences play no role in the evaluation of this set of arguments.

Proposition 3. *If a set of arguments* $\mathcal{S} \in \mathcal{A}$ *is objectively acceptable in all preferred extensions* $\mathcal{E}_{Ags}$ *of* $\mathcal{T}_{Ags}$ *for all the participants Ags in a dialogue, then the set* $\mathcal{S}$ *is the set of most trusted arguments in the dialogue.*

Proof. Since every $\mathcal{E}_{Ag_i}$ is conflict free as the preferred extensions of *PAF* and corresponding F are conflict free, it follows that in $\mathcal{T}_{Ag_i}$, every $\phi_1 \in \mathcal{E}_{Ag_i}$ is either unattacked or attacked by some argument $\phi_2 \in \mathcal{A} \backslash \mathcal{E}_{Ag_i}$ such that $\phi_1 \succ^{\mathcal{A}}_{Ag_i} \phi_2$. For the latter, we know that such attack is *critical* and is *repaired* such that $(\phi_2, \phi_1) \in \mathcal{R}$ becomes $(\phi_1, \phi_2) \in \mathcal{R}$. If ϕ_1 is objectively acceptable in all preferred extensions $\mathcal{E}_{Ags}$ of $\mathcal{T}_{Ags}$, it follows that in all $\mathcal{T}_{Ag_i} \subseteq \mathcal{T}_{Ags}$, ϕ_1 is either unattacked or is attacked by some less preferred argument ϕ_2. Since, $\phi_1 \succ^{\mathcal{A}}_{Ag_i} \phi_2$ denotes that ϕ_1 is more trusted (more preferred) than ϕ_2, it follows that the set of arguments $\mathcal{S} \subseteq \mathcal{E}_{Ags} = \{\phi_1 | \nexists \phi_2 \in \mathcal{A} \backslash \mathcal{E}_{Ags}$ such that $(\phi_2, \phi_1) \in \mathcal{R}$ and $\phi_1 \succ^{\mathcal{A}}_{Ag_i} \phi_2\}$ is the set of most trusted arguments.

6 Example

To illustrate how a participant updates its trust relation with regards to other participants, we provide an extended example, adapted from [16]. We connect the arguments in the dialogue to the participants that advance them as shown in the *Speech Acts* column of Table 1. The *Moves* column of the table shows that the dialogue has two iterations with five moves in the first iteration and four moves in the second iteration. Figures 2 and 3 show the argumentation frameworks derived

from the dialogue by one of the participants Ag_k. $\mathcal{T}^t_{Ag_k}$ represents argumentation framework of Ag_k at iteration t where nodes are arguments and edges are attack relation. Let us consider that participant Ag_k evaluates $\mathcal{T}^1_{Ag_k}$ and $\mathcal{T}^2_{Ag_k}$.

1^{st} ***Iteration***: **Trust Update Rules** $TR^1_{Ag_k}$—In this iteration, Ag_k observes two trust update rules SC w.r.t Ag_i and LJ w.r.t Ag_j. Ag_k observes contradiction in the commitment store of Ag_i (i.e., ϕ_4 attacks ϕ_1 by defending ϕ_2 that attacks ϕ_1). Furthermore, Ag_k observes that Ag_j lacks justification for ϕ_2 as ϕ_5 defeats ϕ_2 (ϕ_2 is defeated by an undefeated argument ϕ_5). Note that the symmetric attack between ϕ_3 and ϕ_4 is obtained by the attack from ϕ_4 to ϕ_3 exchanged via the contradict move, while the ϕ_3 to ϕ_4 attack is known by Ag_k from its knowledge base $\mathcal{KB}_{Ag_k}$.

Preference on Trust Update Rules $\succeq^1_{Ag_{i(TR)}}$—Let Ag_k's preference on the trust update rules be $LJ \succ^1_{Ag_{k(TR)}} SC \succ^1_{Ag_{k(TR)}} AR$.

Trust Update $\succeq^1_{Ag_k}$—Given the trust update rules and Ag_k's preference on the rules, from Definition 15, we can infer that Ag_k prefers (i.e., trusts) Ag_i to Ag_j. Likewise, Ag_k prefers itself to Ag_i (i.e., $\succeq^1_{Ag_k} = Ag_k \succ^1_{Ag_k} Ag_i \succ^1_{Ag_k} Ag_j$).

Ag_k's Conclusion $\mathcal{E}_{Ag_k}$—In $\mathcal{T}^1_{Ag_k}$, Ag_k considers that ϕ_1 and ϕ_5 defeat ϕ_2, ϕ_3 defeats ϕ_4, and $\mathcal{E}^1_{Ag_k}$ is $\{\phi_1, \phi_3, \phi_5\}$.

Table 1. Example: Dialogue

Moves	Speech acts	Arguments
m^1_1	$assert(Ag_i, \phi_1, 1)$	ϕ_1: Death penalty is a legitimate form of punishment
m^1_2	$contradict(Ag_j, \phi_1, \phi_2, 1)$	ϕ_2: God does not want us to kill
m^1_3	$contradict(Ag_k, \phi_2, \phi_3, 1)$	ϕ_3: God does not exist
m^1_4	$contradict(Ag_i, \phi_3, \phi_4, 1)$	ϕ_4: Some people believe in God
m^1_5	$contradict(Ag_k, \phi_2, \phi_5, 1)$	ϕ_5: The legal status of the death penalty should not depend on some random people's belief
m^2_1	$assert(Ag_j, \phi_6, 2)$	ϕ_6: The state has no right to put its subjects to death
m^2_2	$contradict(Ag_i, \phi_6, \phi_7, 2)$	ϕ_7: If child rapists and murderers are put to death it will reduce the number of suicides by the survivors
m^2_3	$contradict(Ag_k, \phi_6, \phi_8, 2)$	ϕ_8: Majority opinion in some democratic countries favour death penalty
m^2_4	$contradict(Ag_j, \phi_7, \phi_9, 2)$	ϕ_9: There is no strong evidence that the death penalty makes victims of child abuse feel good

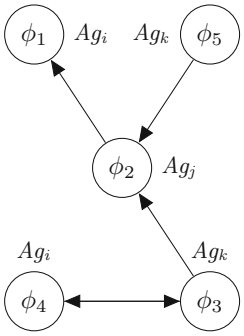

Fig. 2. $\mathcal{T}^1_{Ag_k}$ for 1^{st} iteration

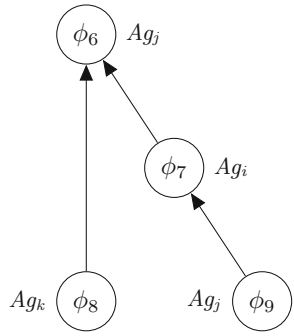

Fig. 3. $\mathcal{T}^2_{Ag_k}$ for 2^{nd} iteration

2^{nd} ***Iteration***: **Trust Update Rules** $TR^2_{Ag_k}$—Ag_k observes that Ag_j lacks justification for ϕ_6 and Ag_i lacks justification for ϕ_7. Therefore, Ag_k observes one trust update rule LJ w.r.t to both Ag_i and Ag_j.
Preference on Trust Update Rules $\succeq^2_{Ag_{k(TR)}}$—Ag_k observes just one trust update rule. Therefore, preference over the trust update rules is not applicable in this iteration.
Trust Update $\succeq^2_{Ag_k}$—Note that, Ag_j has an undefeated argument ϕ_9 in this iteration while Ag_i has none. Therefore, Ag_k prefers Ag_j to Ag_i and itself to Ag_j (i.e., $\succeq^2_{Ag_k} = Ag_k \succ^2_{Ag_k} Ag_i \succ^2_{Ag_k} Ag_j$).
Ag_k's Conclusion $\mathcal{E}_{Ag_k}$—In $\mathcal{T}^2_{Ag_k}$, Ag_k considers that ϕ_8 defeats ϕ_6, ϕ_9 defeats ϕ_7, and $\mathcal{E}^2_{Ag_k}$ is $\{\phi_8, \phi_9\}$.

This example demonstrates how trust evolves in a dialogue and how such trust is used as a basis for expressing preferences between the arguments exchanged in the dialogue. In addition, the example illustrates how trust affects the justified conclusions obtained from a dialogue.

7 Related Work

Recent works on the integration of trust and argumentation has provided paradigms for handling inherent uncertainties in the interactions among agents in multi-agent systems. The importance of relating trust and argumentation was highlighted in [6]. In [10], arguments are considered as a separate source of information for trust computation.

There are four works in the literature which are closely related to the research described in this paper. The first is [12], where the authors propose a model of argumentation where arguments are related to their sources and a degree of acceptability is computed on the basis of the trustworthiness degree of the sources. The model also provides a feedback such that the final quality of the arguments influences the source evaluation as well. In this approach, different dimensions of trust are represented as graded beliefs ranging between 0 and 1

which change across different domains and arguments evaluated by a labelling algorithm. The labelling algorithm computes a fuzzy set of accepted arguments whose membership assigns to each argument a degree of acceptability unlike the extension-based semantics that we apply in our approach.

While related, the work of [12] differs from the current paper in several ways. First, the approach does not consider the cumulative effect of converging sources on argument acceptability. We consider this effect in our model by categorising accepted arguments into two categories namely *objectively acceptable* and *subjectively acceptable* extensions, based on the number of sources that have the arguments acceptable in their extensions. Second, unlike our approach, the evaluation of the trustworthiness degree of a target agent is not induced by the trusting agent's argumentation framework, but determined by the internal mechanism of the trusting agent. Third, [12] considers that in a dialogue, the final acceptability value of the arguments provides a feedback on the trustworthiness degree in the information source. In our approach, we observe that trust can change during the dialogue itself and as such the trust rating of a target participant should be updated at every stage (iteration) of a dialogue.

The works in [15,17] are closely related to ours. The authors present a framework which considers the source of arguments, and expresses a degree of trust in them. They define trust-extended argumentation graphs in which each premise, inference rule and conclusion of an argument is associated with the trustworthiness degree of the source proposing it. In this approach, the trust rating associated with the arguments and their sources does not change. In our approach, trust ratings associated with arguments and sources change between iterations. This notion of dynamic trust rating is captured by socio-cognitive models of trust [4] and other computational trust approaches [5,8].

Lastly, [18] models the connection between arguments about the trustworthiness of information sources and the arguments from the sources—as well as the attacks between the arguments. An information source is introduced into an argumentation framework as a meta-argument and an attack on the trustworthiness of the source is modelled as an attack on the meta-argument. A source is considered trustworthy if its meta-argument is accepted. Like us, [18] model the feedback from sources to arguments and vice-versa. However, like [12], they do not consider how trust evolves in the course of a dialogue.

8 Conclusions

This paper describes how trust changes during argumentation-based dialogues and how such change affects the justified conclusion of the dialogue. In particular, as arguments are exchanged in a dialogue, we formalise a number of trust update rules that a given participant can take into consideration for updating its trust relation over other target participants. The first contribution of our approach is that it captures how trust is dynamically updated in dialectical argumentation and how trust can affect the set of justified conclusions.

It is worth mentioning that the semantics of abstract argumentation frameworks have only focused on identifying which points of view are defensible and

preference-based argumentation frameworks have extended these semantics to deal with preferences between arguments. However, they do not describe *why* one argument should be preferred over another. In our approach, the trust rating of the sources of arguments provides such a basis.

As future work, we intend to find out how change in trust in dialectical argumentation can affect the goals and argumentative strategies of participants. In addition, change in trust during a dialogue may require less trusted participants to present more evidence for their arguments to be believed, while the burden of proof reduces on more trusted participants. This is also an issue for future work. Finally, we are investigating an orthogonal approach to modelling changes in trust within an ongoing dialogue through the use of meta-argumentation. Doing so will eliminate the need for discrete iterations as used in the current work, and an empirical evaluation of the two approaches with regard to human intuitions will allow us to determine which approach is more realistic and useful.

References

1. Amgoud, L., Vesic, S.: Rich preference-based argumentation frameworks. Int. J. Approx. Reason. **55**(2), 585–606 (2014)
2. Bench-Capon, T.J.: Persuasion in practical argument using value-based argumentation frameworks. J. Log. Comput. **13**(3), 429–448 (2003)
3. Bonzon, E., Delobelle, J., Konieczny, S., Maudet, N.: A comparative study of ranking-based semantics for abstract argumentation. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, pp. 914–920 (2016)
4. Castelfranchi, C., Falcone, R.: Trust Theory: A Socio-Cognitive and Computational Model, vol. 18. Wiley, Hoboken (2010)
5. da Costa Pereira, C., Tettamanzi, A.G., Villata, S.: Changing ones mind: erase or rewind? Possibilistic belief revision with fuzzy argumentation based on trust. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (2011)
6. Dix, J., Parsons, S., Prakken, H., Simari, G.: Research challenges for argumentation. Comput. Sci.-Res. Dev. **23**(1), 27–34 (2009)
7. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell. **77**(2), 321–357 (1995)
8. Fullam, K.K., Barber, K.S.: Dynamically learning sources of trust information: experience vs. reputation. In: Proceedings of the 6th International Conference on Autonomous Agents and Multiagent Systems, pp. 164:1–164:8 (2007)
9. Jøsang, A., Keser, C., Dimitrakos, T.: Can we manage trust? In: Herrmann, P., Issarny, V., Shiu, S. (eds.) iTrust 2005. LNCS, vol. 3477, pp. 93–107. Springer, Heidelberg (2005). https://doi.org/10.1007/11429760_7
10. Matt, P.A., Morge, M., Toni, F.: Combining statistics and arguments to compute trust. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, pp. 209–216 (2010)
11. McBurney, P., Parsons, S.: Games that agents play: a formal framework for dialogues between autonomous agents. J. Log. Lang. Inf. **11**(3), 315–334 (2002)
12. Paglieri, F., Castelfranchi, C., Pereira, C.D.C., Falcone, R., Tettamanzi, A., Villata, S.: Trusting the messenger because of the message: feedback dynamics from

information quality to source evaluation. Comput. Math. Organ. Theory **20**(2), 176 (2014)
13. Panisson, A.R., Meneguzzi, F., Vieira, R., Bordini, R.H.: Towards practical argumentation-based dialogues in multi-agent systems. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 2, pp. 151–158. IEEE (2015)
14. Parsons, S., McBurney, P., Sklar, E., Wooldridge, M.: On the relevance of utterances in formal inter-agent dialogues. In: Proceedings of the 6th International Conference on Autonomous Agents and Multiagent Systems, pp. 240:1–240:8 (2007)
15. Parsons, S., Tang, Y., Sklar, E., McBurney, P., Cai, K.: Argumentation-based reasoning in agents with varying degrees of trust. In: Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems, pp. 879–886 (2011)
16. Spanring, C.: Conflicts in abstract argumentation. Cardiff Argumentation Forum (2016)
17. Tang, Y., Cai, K., Sklar, E., McBurney, P., Parsons, S.: A system of argumentation for reasoning about trust. In: Proceedings of the 8th European Workshop on Multi-Agent Systems, Paris, France (2010)
18. Villata, S., Boella, G., Gabbay, D.M., Van Der Torre, L.: A socio-cognitive model of trust using argumentation theory. Int. J. Approx. Reason. **54**(4), 541–559 (2013)
19. Walton, D., Krabbe, E.C.: Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning. SUNY Press, Albany (1995)