Argument and Computation, 2014
Vol. 5, Nos. 2–3, 123–138, http://dx.doi.org/10.1080/19462166.2014.885083

Taylor & Francis
Taylor & Francis Group

Trust and argumentation in multi-agent systems

Andrew Koster*

Department of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

(*Received 15 July 2013; final version received 13 January 2014*)

This survey is the first to review the combination of computational trust and argumentation. The combination of the two approaches seems like a natural match, with the two areas tackling different aspects of reasoning in an uncertain, social environment. We discuss the different areas of research and describe the approaches taken so far, analysing both how they address the problems and the challenges that are unaddressed.

Keywords: argumentation; trust; survey; multi-agent system

1. Introduction

Agreement Technologies is a newly emerging field of research that is concerned with the theory and practice of computer systems that can make agreements in situations where the preferences and beliefs of the participants are different (Ossowski, 2013). In particular, the vision of Agreement Technologies is to study and develop multi-agent systems with the interactions governed by agreements. Both argumentation and trust are seen as vital components for reaching agreements; and both areas deal with conflicts and uncertainty, in a symbiotic manner. In this paper we will survey the recent trends in these areas, particularly highlighting the insights that considering trust brings to the field of argumentation, and vice versa, the contributions of argumentation to the field of trust.

The study of argumentation has been approached from two different perspectives. The first approach is to consider argumentation as a model for reasoning. The formalisation of argumentation as a non-monotonic logic has been studied by Pollock (1987), Dung (1995) and many others. Philosophers, behavioural scientists, psychologists and sociologists, such as Walton (1990) and Mercier and Sperber (2011), study the relationship between argumentation and human reasoning. The second approach, is to consider the dialectical theory of argumentation, and in particular the role of argumentation in communication. This approach is taken by, among others, Grice (1975), Walton and Krabbe (1995) and Prakken (2006).

This dualistic approach to argumentation has carried over its use in multi-agent systems (Rahwan, 2006): argumentation is studied, and used, to model agent reasoning and structure communication between agents. Argumentation has been used in belief revision for intelligent agents (Paglieri & Castelfranchi, 2006) and collaborative, multi-agent planning (Pardo, Pajares, Onaindia, Godo, & Dellunde, 2011). For communication between agents, argumentation has played a crucial role in specifying dialogue protocols.

We distinguish a similar split in the approaches to using trust and argumentation together. In agent reasoning, trust in information sources has been used in argumentative reasoning. Moreover, argumentation has been used to reason about trust evaluations. In multi-agent interaction, argumentation is used to communicate about subjective trust evaluations, and trust is used for deciding

*Email: koster.andrew@gmail.com

on guiding whom to communicate with. Finally, argumentation can be used to communicate between the computational trust model and a human user of the system.

This work aims at surveying the state of the art in combining trust and argumentation. Because of the diverse aims of the different lines of research we discuss, many of these works cannot be directly compared. Nevertheless, we feel there are many similarities in the approaches taken, because they all aim to improve an agent's ability to cope in an uncertain environment. We distinguish the works along two principal axes:

(1) Whether the primary aim of the work is to improve a single agent's reasoning, or to enhance communication between agents.
(2) Whether computational trust is used within an argumentation framework, or whether argumentation is used in a computational trust model.

Works that deal primarily with agent reasoning are discussed in Section 2, which in turn is split into Section 2.1, about using trust to improve the argumentation framework, and Section 2.2, about using argumentation in a trust model. In Section 3 we discuss the works that deal primarily with communication. Finally we summarise our findings in Section 4.

2. Trust, argumentation and reason

Trust and argumentation have been used in agent reasoning in two principal areas. The first is in belief revision. Belief revision is the problem of keeping an agent's beliefs updated in the face of incomplete and conflicting information (Alchourrón, Gärdenfors, & Makinson, 1985).

Recently, argumentation has been studied as providing a mechanism for performing belief revision (Paglieri & Castelfranchi, 2006). One of the problems in belief revision is to decide what should be believed when conflicting information is received from different sources. Argumentation can help with this by providing a computationally efficient framework for selecting admissible sets of arguments. Another approach is to discard information from less credible (or trustworthy) sources (Barber & Kim, 2001). These two approaches can be combined, by seeing the trustworthiness of sources as meta-level arguments about the admissibility of the provided information. We will discuss such approaches in further detail in the next section.

2.1. *Belief revision using argumentation and trust*

The traditional approach to belief revision is *prioritised* belief revision, in which newly received information is always given priority over the agent's currently held beliefs. However, this does not take into account that new information may come from an unreliable source; regardless of whether this is due to a faulty sensor or communication with another agent. In particular, other agents may be unreliable for any number of reasons, e.g. they are intentionally lying, their own sensors are unreliable, or they use different subjective criteria to evaluate a situation. In a dynamic, social environment, it thus seems necessary to drop (some of) the AGM postulates (Alchourrón et al., 1985), which define a rigorous set of conditions for any belief revision operator, in favour of some other system.

Barber and Kim (2001) propose a method that takes the reputation of information sources into account in a probabilistic belief revision method. Their method works on a belief base, and uses a non-prioritised belief revision operator of the type 'expansion + consolidation' (Hansson, 1999): they first add all the new information to the belief base set-theoretically, and then delete beliefs from this set, according to some criteria, to make the belief base consistent. Barber and Kim aim to keep their set of beliefs maximally certain. The certainty of an individual belief is calculated using a Bayesian method that combines the reputation of each information source with the certainty

values the source communicated about the belief. However, such a method only works on belief bases, and makes no distinction between the new information and old beliefs. Particularly if the trustworthiness of information is equal there are good arguments for accepting new information, especially in dynamic environments.

2.1.1. *Fuzzy argumentation and trust*

da Costa Pereira, Tettamanzi, and Villata (2011) define a belief revision operator that works along similar lines to Barber and Kim's: it also follows the 'expansion + consolidation' method, however the consolidation step is very different, and allows conflicting beliefs to persist, in a way, so that this information is not lost, if a later revision reinstates some beliefs. This is achieved through the use of fuzzy argumentation. An agent's belief base is a possibilistic belief base, as described by da Costa Pereira and Tettamanzi (2010), however, underlying that belief base is a fuzzy argumentation framework. The degree to which a formula is believed in the belief base depends directly on the degree to which the arguments supporting and attacking it are acceptable.

If the acceptability of arguments attacking the formula are stronger than the support, the formula is not believed, and the degree of belief is 0. If the support is stronger than the attacks, then the degree of belief is the difference between the two. We will describe the details of the fuzzy argumentation framework below, but it is important to note that this allows inconsistent arguments to coexist in the argumentation framework: they are simply arguments for and against a formula, while the belief base is guaranteed to be consistent. New information changes the labelling in the argumentation framework, which can cause an entirely different set of formulas to be accepted in the belief base, but information is never erased: it persists in the underlying argumentation framework, and changing trustworthiness, or the reception of new evidence can cause beliefs to be reinstated. In contrast, using Barber and Kim's revision operator, or in most prioritised revision operators, when the belief base is revised with information that is inconsistent with the current base, beliefs are irrevocably erased to be consistent with the new information.

The fuzzy argumentation framework that da Costa Pereira et al. use, is an extension of the reinstatement labelling framework, proposed by Caminada (2006). An argument in the framework is a *deductive argument* (Besnard & Hunter, 2001): an argument is a tuple $\langle \Psi, \psi \rangle$, such that Ψ is consistent, and a minimal proof for entailing ψ, in a propositional language $\mathcal{L}$. An argument can attack another argument, either by *undercutting* it, or *rebutting* it. $\langle \Phi, \phi \rangle$ undercuts $\langle \Psi, \psi \rangle$ iff $\phi = \neg(\psi_1 \wedge \cdots \psi_n)$ and $\{\psi_1, \ldots, \psi_n\} \subseteq \Psi$. $\langle \Phi, \phi \rangle$ rebuts $\langle \Psi, \psi \rangle$ iff $\phi \Leftrightarrow \neg\psi$ is a tautology.

Da Costa Pereira et al. introduce a fuzzy labelling system over the arguments by considering the trustworthiness of the sources. In Caminada's framework, an argument is labelled as either *in, out* or *undecided*. Da Costa Pereira et al. generalise this by considering the acceptability of an argument as a degree between 0 and 1. They first introduce the trustworthiness of an argument, by defining it as the trustworthiness of the most trustworthy source that proposes it. The acceptability of an argument is then computed in an iterative process. At time $t = 0$ they define α_0 of an argument A as its trustworthiness $\text{trust}(A)$. At each subsequent time, the acceptability of an argument A is defined as:

$$\alpha_{t+1} = \frac{1}{2}\alpha_t(A) + \frac{1}{2}\min\{\text{trust}(A), 1 - \max_{B \in \mathbf{B}}\{\alpha_t(B)\}\},$$

where $\mathbf{B}$ is the set of arguments that attack A.

This is based upon the intuition that the acceptability of an argument cannot be larger than its trustworthiness, and secondly that it can be lowered if arguments that attack it are sufficiently trustworthy. They prove this iterative process converges, and in general use the acceptability labelling $\alpha(A) = \lim_{t \to \infty} \alpha_t(A)$ for all arguments A. The speed of this convergence is linear in the number of arguments.

As stated above, the acceptability of arguments is then used to generate a possibilistic belief base and for details we refer to da Costa Pereira et al. (2011). However, from the way the acceptability is computed, it is immediately obvious that new information can have one of two effects: the first is that it adds a new argument, which attacks one of the arguments in the agent's belief base. The second is that new information causes the trustworthiness of arguments to change (for instance, by updating the trustworthiness of the sources, or because a new source adds its support to some argument). Both of these cause belief base revision by triggering the iterative process that computes the acceptability of the arguments, which in turn causes the belief base to change.

Nevertheless, this approach has the same disadvantages as other 'expansion + consolidation' approaches: it deals only with belief base revision, not belief set, and the priority of new information is not explicitly taken into account. In the next section we discuss another approach that avoids these issues. Additionally, the method does not account for the number of sources that support an argument: an argument's acceptability is defined as the maximum trustworthiness of any source supporting it. This could be augmented by more advanced trust models.

2.1.2. *Trust-based selection in belief revision*

Tamargo, García, Thimm, and Krümpelmann (2012) propose an alternative method for using argumentation to perform the belief revision that defines a non-prioritised belief revision operator of the type 'decision + revision'. They first decide what part, if any, of the new information should be accepted, and then use a prioritised revision operator to revise the agent's beliefs with this accepted information. This method works on belief sets, and explicitly takes the priority of new information into account, although it can be discarded if the source is not sufficiently credible. Similar to da Costa Pereira et al. and unlike Barber and Kim, Tamargo et al. do not attempt to model the trustworthiness of sources, but assume there is a mechanism in place to rate the credibility of sources. Differently to either of the previous approaches, however, they assume that each agent has a total preorder $\leq$ on the set of all agents in the system. Each belief is annotated with its source. The credibility ordering over the sources can then be used to define a preference ranking between arguments.

Tamargo et al. define, just as da Costa Pereira et al. do, an argument as a deductive argument. The credibility of the sources is then used to define a preference ranking between arguments. An argument $\langle \Psi, \psi \rangle$ is at least as preferred, to agent A, as another $\langle \Phi, \phi \rangle$ if and only if for all $(B : \psi) \in \Psi$ there is a $(C : \phi') \in \Phi$ such that $C \leq_A B$. In other words, the least credible source in the argument is at least as credible as the least credible source in the other argument.

This preference ranking over arguments is used in defining a belief revision operator. This is done, additionally, using argument trees. An argument tree is a tree where the nodes are arguments. The children of a node are arguments that attack the parent. Tamargo et al. do not consider rebuttal attacks, although they could easily be incorporated, but only consider undercutting attacks between arguments, which are defined as described above.

Whether or not an incoming piece of information is accepted is decided using a *categoriser*, which assigns a value to an argument tree, depending on how strongly this argument tree favours the root argument. For an agent A, they define this categoriser γ as:

$$\gamma_A(\tau) = 1 - \max(\{\gamma_A(\tau') \mid \tau' \in \text{children}(\text{root}(\tau)) \wedge \text{root}(\tau) \preceq_A \text{root}(\tau')\}),$$

where τ is an argument tree, $\text{root}(\tau)$ is the root argument of an argument tree and $\text{children}(N)$ gives the subtrees of node N. If a tree consists of a single node, it is categorised with value 1. An argument tree is thus categorised as 0 if there is at least one sub-tree under the root that is categorised with value 1: in other words, there is at least one child argument that is preferred over the root, but has no subtrees of its own that defeat it.

This credibility categoriser is then used to evaluate new information. Let agent A have beliefs $\mathcal{B}_A$ and receive a set of new information $\mathcal{I}$ from agent A_i. Now, for any $\phi \in \mathcal{I}$, we can do the following: the set $\mathcal{P}_\phi$ is the set of argument trees for ϕ that can be constructed in $\mathcal{B}_A \cup \mathcal{I}$, and $\mathcal{C}_\phi$ is a similar set of argument trees for $\neg\phi$. κ is a simple accumulator and is defined as follows:

$$\kappa(\mathcal{P}_\phi, \mathcal{C}_\phi) = \sum_{\tau \in \mathcal{P}_\phi} \gamma_A(\tau) - \sum_{\tau \in \mathcal{C}_\phi} \gamma_A(\tau).$$

Agent A *credulously accepts* ϕ iff $\kappa(\mathcal{P}_\phi, \mathcal{C}_\phi) \geq 0$. Or, in other words, there are at least as many reasons to believe in ϕ as there are to believe in $\neg\phi$. The agent *skeptically accepts* ϕ iff $\kappa(\mathcal{P}_\phi, \mathcal{C}_\phi) > 0$. These two methods define two different ways that allow the agent to decide what information in $\mathcal{I}$ to accept, and what to reject. Tamargo et al. show how to use this with a prioritised belief revision function in order to revise the agent's beliefs with new information. They then prove that this revision operator satisfies a number of desiderata for revision operators.

Nevertheless, this work leaves a number of open questions. The first is regarding the computational complexity. An argument in this system is a minimal deductive proof for a formula. However, for propositional logic, the problem of finding a single proof is an NP-complete problem, and the method requires all such proofs for every formula, and its negation, in the set of new information. This is in addition to the computational cost of the prioritised belief revision, which is, in and of itself, NP-complete. While no implementation details were given, this seems prohibitive for anything except the smallest of belief sets.

Secondly, it requires the credibility ordering over the different agents to be a total preorder. In many environments it may not be realistic to expect to have information about the trustworthiness of every potential source. This could probably be solved in a pragmatic manner, because the credibility ordering is only required upon receiving new information from a source, at which point the agent could endeavour to learn how trustworthy that source is, using conventional methods. Nevertheless, they do not go into the details of computational trust models for this purpose, which can be highly context-dependent, and have to deal with the uncertainty and dynamicity of the environment themselves.

Tang, Cai, McBurney, Sklar, and Parsons (2012) take a very similar approach to belief revision as Tamargo et al. do, although they do not formalise their framework in terms of belief revision, but see it rather as an abstract argumentation framework that is augmented with trust. The two approaches are largely complementary, with Tang et al. taking a more pragmatic approach and Tamargo et al. proving that the system satisfies some desirable properties.

Tang et al. use a very similar formalisation of arguments to Tamargo et al. (despite some representational differences), but the main difference between the two methods is that Tang et al. use a so-called trust network. A trust network is a directed graph, in which each node is an agent, and every arc is a trusting relationship, which is labelled with the trust evaluation. An important restriction that they require is that trust is transitive. In other words, if A trusts B, and B trusts C, then A also trusts C: the evaluation, or quantification, of that relationship is left uninstantiated in the general model. In their example scenarios, the evaluation is numerical, and they take the minimum of the values to aggregate a trust evaluation over a path in the network.

By assuming transitivity of trust, they solve one of the issues of Tamargo et al.'s framework: namely that the credibility ordering is a total preorder over the agents. Because of the structure of the agent-centric trust network, each agent has an associated trust evaluation, resulting in exactly such a total preordering.

Tang et al., however, do not specify where the arguments come from. Tamargo et al. require all arguments, or proofs, for a formula to be evaluated, in order to decide whether or not it should be accepted. This seems necessary in order to define a belief revision operator, but Tang et al. leave this

question open, and take the more pragmatic approach of specifying how to incorporate trust into an abstract argumentation framework. Thus, rather than proving properties about a belief revision operator, they discuss what impact different methods for aggregating transitive trust has upon the credibility of the arguments. In particular, Parsons, Tang, Sklar, McBurney, and Cai (2011), describe a number of properties that the transitive trust operator can have, and consequently prove whether or not certain desiderata of the argumentation system hold.

Nevertheless, it is unclear whether the properties described, or transitivity of trust itself, actually hold in any real system. A number of prominent computational trust models, such as TidalTrust (Golbeck, 2006) or the model presented by Yu and Singh (2002), rely on transitivity, and are used successfully. However, assuming trust is transitive can probably only be done in certain domains and does not hold in the general case (Falcone & Castelfranchi, 2012). This must be taken into account when using Tang et al.'s framework.

2.2. *Argumentation-based trust*

In the previous section we discussed the state-of-the-art developments in using trust to aid in argumentation-based reasoning. A simultaneous development is that argumentation can be used to reason about trust itself. In particular, trust is treated as an integral part of the reasoning process, following a socio-cognitive interpretation of it (Castelfranchi & Falcone, 2010). Many of the computational trust models tackle the problem of how to compute an estimation of a target's trustworthiness, based on that target's past behaviour, its reputation, its social connections, etc. (Jøsang, Ismail, & Boyd, 2007). This estimate must be incorporated into a decision process, in which an agent's trust in another is a consciously deliberated reliance on it, for a specific purpuse. A trust evaluation must thus be considered as a mental attitude, that the agent reasons about.

There are a number of approaches that explicitly deal with reasoning about trust, but as argued by Pinyol and Sabater-Mir (2013), incorporating trust into a computational model with explicitly represented mental states is a recent development. Prade (2007) was one of the first to propose a computational framework for reasoning about trust, and states that the reason for using an argumentative framework is that: "a trust assessment is the result of a decision process, based on arguments in favour or against a classification of the agent or the source to be evaluated, at a particular trust level". In other words, argumentative reasoning is capable of dealing with the various different reasons for trusting, and not trusting, the target. Moreover, as Stranders, de Weerdt, and Witteveen (2008) point out, argumentative reasoning is easy to understand for human users of the system, so it can be used to explain *why* a particular agent is trusted, or not.

2.2.1. *Abductive argumentation for evaluating trust*

Prade (2007) explicitly models uncertainty in his representation of trust: a trust evaluation is a range $[t^-, t^+]$, with both t^- and t^+ in some range of possible trust values S. The meaning of this is that the agent knows that the true trustworthiness τ is no smaller than t^- and no greater than t^+. If $t^- = t^+$ then the value is crisp.

The trust model itself consists primarily of a rule-based system K, which relates levels of trust to observable behaviours of the target. The rules in this system are abductive arguments of one of two forms: let $\rho, \sigma \in S$ and φ a literal in some language for describing observable behaviour, then a rule has the form:

- 'if $t^- \geq \rho$ then φ'
- 'if $t^+ \leq \sigma$ then φ'

After observing the target's actual behaviour represented as a set of literals (the observations), an agent can abduce what t^- and t^+ must be to optimise consistency with the rules in K. If this results in $t^+ < t^-$, which is not ruled out by the structure of the rules, then the agent should be cautious (or pessimistic), and choose $t^+ = t^-$.

One thing that sets this work apart from most of the other work that argues about trust is that the argumentation is performed in the opposite direction from the one expected: Prade formulates the arguments such that the trust evaluation appears in the support, rather than the conclusion of the arguments. This is why the evaluation of trust is done using abductive reasoning, rather than deductive. It also does not use a full argumentation framework, in which different arguments attack each other. In contrast, Stranders et al. (2008) do allow for this, and we will see that this is taken even further in the framework of Villata, Boella, Gabbay, and van der Torre (2013).

2.2.2. *Fuzzy argumentation for explaining trust*

Stranders et al. (2008) present a decision-support system in which an agent recommends an interaction with another agent: in other words, it recommends whom to trust. The system uses argumentation, based on the framework presented by Amgoud and Prade (2004), to provide the reasoning underlying that decision.

The first step in the decision process is to induce an opponent model from historical data. For this purpose Stranders et al. use a Fuzzy Rule Learner to learn a base of fuzzy rules that describe agents' expected behaviour. These rules are formulas in a fuzzy logic with the structure 'if a is A then c is C', where a is some attribute (such as the capability or honesty of the target) to be learned over, c is the variable to be learned (such as the quality of a delivered product) and A and C are fuzzy sets. They further define two measures to quantify how well the rule fits the data, *confidence* and *match strength*.

This rule base is used to construct an argument for deciding whether or not to trust a target agent, using a slightly modified version of Amgoud and Prade's fuzzy argumentation framework. An argument is a triple $A = \langle S, C, d \rangle$, with S the support of the argument, C the consequences and d a decision, such that $S \cup \{d\}$ is a minimal proof for C. The formulas in S are elements of the agent's knowledge base, which contains the fuzzy rules that were learned, and the agent's observations. The formulas in C are goals of the agent, with an associated *weight*. Using the confidence and match strength of the rules in S, and the weight of the goals in C, Stranders et al. define a way of computing the *strength* of an argument. One decision is preferred over another if the argument supporting the former is stronger than the argument supporting the latter.

Thus arguments for and against trusting a user can be formulated, for instance: let there be rules 'if capability is low then quality is low' and 'if willingness is high then delivery time is good'. An argument for the decision to trust a target might be based on an observation that the target's willingness is high, while simultaneously an argument for the decision not to trust a target can be based on the observation that he is incapable. The comparative weight of receiving a quality product and a good delivery time play a role in deciding what to recommend to the user, as well as how well the rules are estimated to fit the environment.

The argumentative structure allows this to be presented in a way that is understandable to a user. The main computation of this framework is in inducing fuzzy rules. The example scenario in which they test the framework is very restricted, and the induction process is straightforward. However, generally speaking, the induction of predictive theories from examples, is undecidable; even the more restricted practical tasks are known to be not polynomially PAC-learnable. It is thus not clear how well this method for inducing arguments from data works for more realistic scenarios. Matt, Morge, and Toni (2010) sidestep this computational problem by creating a hybrid system,

using a Dempster-Shafer belief function to analyse the target's past behaviour, and argumentation to augment this with justified claims about the target's expected behaviour.

Nevertheless, Stranders et al.'s model is interesting precisely because it highlights an often overlooked aspect of computational trust: in most multi-agent systems, the agents exist to support users, who must authorise any behaviour: being able to explain the trust evaluation is important in this context.

2.2.3. *Combining statistics and argumentation*

Matt et al. (2010) build upon an established computational trust model (Yu & Singh, 2002) that uses a Dempster–Shafer belief function to estimate the trustworthiness of a target. They do this by first partitioning the past interactions with a target into three sets: the interactions in which the outcome was good, the interactions in which the outcome was bad, and the interactions in which the outcome was inappreciable (the rest). An agent is deemed trustworthy if the difference between the frequency of good interactions and the frequency of bad interactions is greater than a prespecified threshold.

Simultaneously, they use an abstract argumentation framework (Dung, 1995) to specify two types of arguments: forecast arguments support trusting or distrusting a target, and mitigation arguments attack forecast arguments. These arguments are generated from the contracts between agents. A contract regulates the interactions between two agents and provides guarantees about the quality of these interactions along various dimensions. In particular, they specify *availability, security, privacy* and *reliability*, but it is easy to see that in different domains, different dimensions are important. The arguments for each dimension d are generated as follows:

- An argument forecasting untrustworthy behaviour, based on the fact that the contract does not provide any guarantee regarding d.
- An argument forecasting trustworthy behaviour, based on the fact that there is a contract guaranteeing a suitable quality of service along dimension d.
- An argument that mitigates a forecasting argument of the second type, on the grounds that the target has, in the past, 'most often' violated its contract clauses concerning d.

They consider trust along each of the dimensions separately and incorporate the argument-based and statistical evidence into a single evaluation by defining an argument-based evidence mass function that evaluates all the evidence in a numerical manner, combining the strength of arguments, their informational value and the statistical evidence. By incorporating more information, the agent should be able to obtain more accurate trust evaluations, and Matt et al., show this empirically.

The main advantage of this method is that it provides a clear computational method for integrating argumentation into a trust model. They are deliberately vague on the content of the arguments, and work with an abstract framework. They generate the arguments from underlying contracts, but in principle there is no reason why the framework must be instantiated this way, and it should be possible to consider more complex argumentational structures.

2.2.4. *Meta-argumentation*

Villata et al. (2013) take an entirely different approach, and their model can best be considered a hybrid approach, where the trustworthiness of sources is used to evaluate arguments, and in turn arguments can support or attack the trustworthiness of sources.

Villata et al. argue that what is lacking in the models discussed in Section 2.1, is an explicit representation of the sources themselves. One fundamental aspect of argumentation frameworks

is that an argument is acceptable if it is not attacked. However, if a source of an argument is untrustworthy, the argument might be unacceptable even if there is no conflicting information.

In order to remedy this, they propose an extended argumentation framework (Boella, Gabbay, van der Torre, & Villata, 2009) in which sources are represented explicitly. The argumentation framework they propose can be seen in two ways: the first is with sources, outside of the framework, proposing arguments. The second is a meta-level in which the previous representation is flattened and the sources are represented explicitly within an argumentation framework. At this meta-level there are a number of auxiliary arguments, and for details we refer to (Villata et al., 2013), but in particular there is the argument trust(A), where A is a source. Because the meta-level is a regular argumentation framework, arguments can attack these meta-arguments. In Villata et al.'s example scenario, one agent proposes a rebuttal attack against the trustworthiness of another. This instantiates a second meta-level argument.

Villata et al. formalise this all, and provide the rules for when a meta-level argument may be instantiated in such a way to prevent infinite regression from happening (for instance, if two agents attack the trustworthiness of each other). Dung-style semantics can then be applied to decide which arguments are accepted.

In many ways this framework is similar to the trust-based belief revision method proposed by Tang et al. (that we discussed in Section 2.1.2): they both extend an argumentation framework by explicitly linking the trustworthiness of sources to the arguments they propose. However, Tang et al. only represent trust, and use this to define the strength of the proposed arguments, whereas Villata et al. focus explicitly on distrust: arguments can attack the trustworthiness of a source. Particularly, in Tang et al.'s framework, if an argument is not attacked, it is necessarily accepted, because all sources are necessarily trusted to some degree, due to the transitivity of trust in their trust network. Villata et al. assume a similar starting point: all agents are trusted, unless there are arguments to the contrary, but they provide the structure for explicitly representing these arguments.

3. Communicating trust

In the previous section we discussed approaches that combine trust and argumentation for the purpose of a single agent reasoning in a social context. However, argumentation is a method of reasoning that can easily be performed by multiple agents, working together. In particular, argumentation can be used to convince another agent that some claim is justified.

Whereas Stranders et al. (2008) are principally interested in communicating a trust evaluation to a human user, the two approaches we discuss in this section use argumentation to justify an agent's trust evaluation to one another. The focus is thus less on the argumentation framework used, and more on the language and dialogue protocol, which are integral to the communication of arguments. In particular, the works identify argumentation as a means to justify agents' subjective trust evaluations to one another. The first work, by Pinyol (2011) allows agents to decide whether or not a communicated trust evaluation is sufficiently justified. The second work (Koster, Sabater-Mir, & Schorlemmer, 2012a) is previous work done by the author, which builds upon Pinyol's framework to allow a third party agent to personalise its trust evaluation to the receiver.

3.1. *Arguing about trust*

Pinyol (2011) starts by modelling the trust model as an inference relation between sentences in $\mathcal{L}_{\text{Rep}}$, a first-order language about trust and reputation. This language is defined by a taxonomy of terms used for describing the process of computing trust. A trust model is considered as a computational process: given a finite set of inputs, such as beliefs about direct experiences or reputation, it calculates a trust evaluation for a target. The semantics of a computational process

can be given by the application of a set of inference rules (Jones, 1997). A trust model can thus be represented by a (finite) number of applications of inference rules that allows for the inference of a trust evaluation δ, given some input Δ.

The inference rules themselves depend on the specifics of the computational process and thus the actual trust model being used, but for any computational trust model, such an inference relation exists. For instance, a trust model might have a rule:

$$\frac{\text{img}(T, X), \text{rep}(T, Y)}{\text{trust}(T, (X + Y)/2)}.$$

With img, rep and trust predicate symbols in $\mathcal{L}_{\text{Rep}}$ and T, X and Y variables. For a specific target Jim, an agent knows $\{\text{img}(\text{Jim}, 3), \text{rep}(\text{Jim}, 5)\}$. It can thus infer $\text{trust}(\text{Jim}, 4)$ using the rule above. For a full example of representing a trust model in inference rules, we refer to Pinyol and Sabater-Mir (2009).

3.1.1. *Reasons for having a trust evaluation*

Arguments are sentences in the $\mathcal{L}_{\text{Arg}}$ language. This language is defined over $\mathcal{L}_{\text{Rep}}$. A sentence in $\mathcal{L}_{\text{Arg}}$ is a formula $(\Phi : \alpha)$ with $\alpha \in \mathcal{L}_{\text{Rep}}$ and $\Phi \subseteq \mathcal{L}_{\text{Rep}}$. This definition is based on the framework for defeasible reasoning through argumentation, given by Chesñevar and Simari (2007). Intuitively Φ is the defeasible knowledge required to deduce α. Defeasible knowledge is the knowledge that is rationally compelling, but not deductively valid. The meaning here, is that using the defeasible knowledge Φ and a number of deduction rules, we can deduce α. The defeasible knowledge is introduced in a set of elementary argumentative formulas. These are called *basic declarative units*.

DEFINITION 3.1 Basic Declarative Units: A basic declarative unit (bdu) is a formula $(\{\alpha\} : \alpha) \in \mathcal{L}_{\text{Arg}}$. A finite set of bdus is an argumentative theory.

Arguments are constructed using an argumentative theory Γ and the inference relation $\vdash_{\text{Arg}}$, characterised by the deduction rules Intro-BDU, Intro-AND and Elim-IMP.

DEFINITION 3.2 Deduction rules of $\mathcal{L}_{\text{Arg}}$

$$\text{Intro-BDU:}\frac{}{(\{\alpha\} : \alpha)},$$

$$\text{Intro-AND:}\frac{(\Phi_1 : \alpha_1), \ldots, (\Phi_n : \alpha_n)}{(\bigcup_{i=1}^{n} \Phi_i : \alpha_1 \wedge \cdots \wedge \alpha_n)},$$

$$\text{Elim-IMP:}\frac{(\Phi_1 : \alpha_1 \wedge \cdots \wedge \alpha_n \rightarrow \beta), (\Phi_2 : \alpha_1 \wedge \cdots \wedge \alpha_n)}{(\Phi_1 \cup \Phi_2 : \beta)}.$$

An argument $(\Phi : \alpha)$ is valid on the basis of argumentative theory Γ iff $\Gamma \vdash_{\text{Arg}} (\Phi : \alpha)$. Because the deduction rules, and thus $\vdash_{\text{Arg}}$, are the same for all agents, they can all agree on the validity of such a deduction, however each agent builds its own argumentative theory, using its own trust model. Let $\mathcal{I}$ be the set of inference rules that specify an agent's trust model. Its bdus are generated from a set of $\mathcal{L}_{\text{Rep}}$ sentences Δ as follows:

- For any ground element α in Δ, there is a corresponding bdu $(\{\alpha\} : \alpha)$ in $\mathcal{L}_{\text{Arg}}$.
- For all $\alpha_1, \ldots, \alpha_n$ such that $\Delta \vdash \alpha_k$ for all $k \in [1, n]$, if there exists an applicaton of an inference rule $\iota \in \mathcal{I}$, such that $(\alpha_1, \ldots, \alpha_n)/\beta$, then there is a bdu $(\{\alpha_1 \wedge \cdots \wedge \alpha_n \rightarrow \beta\} : \alpha_1 \wedge \cdots \wedge \alpha_n \rightarrow \beta)$, i.e. there is a bdu for every instantiated inference rule for the trust model specified by $\mathcal{I}$.

Continuing the example from above, our agent might have bdus:

$(\{\text{img}(\text{Jim}, 3)\} : \text{img}(\text{Jim}, 3))$,
$(\{\text{rep}(\text{Jim}, 5)\} : \text{rep}(\text{Jim}, 5))$ and
$(\{\text{img}(\text{Jim}, 3) \wedge \text{rep}(\text{Jim}, 5) \rightarrow \text{trust}(\text{Jim}, 4)\} : \text{img}(\text{Jim}, 3) \wedge \text{rep}(\text{Jim}, 5) \rightarrow \text{trust}(\text{Jim}, 4))$.

These bdus constitute an argumentative theory, from which $(\Phi : \text{trust}(\text{Jim}, 4))$ can be inferred, with Φ the union of the defeasible knowledge of the argumentative theory. Similarly, working backwards, an agent can build a valid argument supporting a trust evaluation it believes. Moreover, it can communicate this argument. The other agent, upon receiving such an argument can decide whether or not to accept the trust evaluation. By doing so, the agent effectively filters out communicated trust evaluations that do not coincide with its own frame of reference.

However, if trust evaluations can be based on many different criteria, agents might reach the point where they filter out too much information. To reduce the amount of information discarded, agents, when sending a trust evaluation, could personalise their trust evaluations to the receiver.

3.2. *Personalised trust recommendations*

The framework for personalised trust recommendations builds upon the argumentation framework we described in the previous section, however it allows agents to communicate about more than just the trust evaluations: it allows agents to connect these trust evaluations to their beliefs and goals. The sender can then tailor its trust model to give a trust recommendation tailored to the receiver's goal, or the two agents can argue about their beliefs about the environment. In this manner agents can personalise their trust recommendations to each other. The framework, as presented by Koster et al. (2012a) uses an abstract view of a trust model that allows for the adaptation of the parameters. It specifies a method for connecting these parameters to the agent's beliefs and goals and the argumentation framework is built upon this.

To do so, an agent must be able to reason about trust. AdapTrust (Koster, Schorlemmer, & Sabater-Mir, 2013) defines a meta-model for reasoning about computational trust. It is an extension of the Beliefs-Desires-Intentions framework for intelligent agents (Rao & Georgeff, 1991), and defines how a trust evaluation is connected to other beliefs an agent has, its desires and also its intentions.

AdapTrust defines two new logical structures. The first is a logical language for describing the importance between various criteria, which prescribe values for the parameters in a trust model. For instance, in the example above, img and rep were given equal importance. However, it is possible to prioritise image over reputation or vice versa, and in this case a weighted average should be used in the computation of trust. The value of this weight is given by the way an agent prioritises between past experiences (img) and communication (rep).

These priorities are connected to the agent's beliefs, desires and intentions through a rule-based system. An agent's beliefs, desires and intentions define the priorities that should be used for adapting the trust model to an agent's current situation. Additionally this is how the multi-faceted aspect of trust is emphasised: the goal the agent is trying to achieve influences the priority system and thus the trust model. For instance, in an eCommerce scenario a vendor may be evaluated with regards to the price he charges, and how long it takes him to deliver the goods. Normally, an agent may prefer cheap vendors, but if an agent needs to buy an item urgently, this could change. This is encoded in a priority rule as follows: $\text{buy\_urgent}(X) \rightsquigarrow (\text{delivery\_time} \succ_w \text{price})$. Consequently, the trust model's parameters are modified to prioritise delivery time over price, and we see that the trust model is adapted to the current goal of the agent.

3.2.1. *Personalising trust recommendations*

The argumentation framework by Pinyol et al. that we described in Section 3.1.1 does not allow us to completely address the question of what criteria play a role in computing a trust evaluation, let alone connect these to underlying beliefs and goals. AdapTrust can answer this, but does not provide a language in which to do so. Koster et al. (2012a) extend Pinyol et al.'s argumentation framework with concepts from AdapTrust. The priorities that define the trust model's parameters can be incorporated into the argumentative theory. For this, the dependency of the trust model on the beliefs and goal of an agent must be represented in $\mathcal{L}_{\text{Arg}}$. Rather than using $\mathcal{L}_{\text{Rep}}$ as the single language on which the argumentation framework is built, the agent can argue about concepts in $\mathcal{L}_{\text{KR}} = \mathcal{L}_{\text{Rep}} \cup \mathcal{L}_{\text{PL}} \cup \mathcal{L}_{\text{Rules}} \cup \mathcal{L}_{\text{Bel}} \cup \mathcal{L}_{\text{Goal}}$, where $\mathcal{L}_{\text{PL}}$ and $\mathcal{L}_{\text{Rules}}$ are the languages of the priorities and priority rules, respectively, in AdapTrust, $\mathcal{L}_{\text{Bel}}$ the language of the agent's beliefs and $\mathcal{L}_{\text{Goal}}$ that of the agent's goals.

This allows us to redefine the set of bdus and thus the argumentative theory in such a way that the argumentation supporting a trust evaluation can be followed all the way down to the agent's beliefs and goal. The deduction rules are the same as in Pinyol et al.'s framework, but the bdus for $\mathcal{L}_{\text{Arg}}$ are defined as follows:

DEFINITION 3.3 Basic Declarative Units for $\mathcal{L}_{\text{Arg}}$ Let $\mathcal{I}_{\Psi,\gamma}$ be the set of inference rules that specifies an agent's trust model, in the context of goal γ and beliefs Ψ.

(1) For any sentence $\psi \in \Psi$, there is a corresponding bdu $(\{\psi\} : \psi)$ in $\mathcal{L}_{\text{Arg}}$.
(2) The goal γ has a corresponding bdu $(\{\gamma\} : \gamma)$ in $\mathcal{L}_{\text{Arg}}$.
(3) For each priority rule $\Phi \rightsquigarrow \pi$ such that $\Phi \subseteq (\Psi \cup \{\gamma\})$, $(\{\Phi \rightarrow \pi\} : \Phi \rightarrow \pi)$ is a bdu of $\mathcal{L}_{\text{Arg}}$.
(4) For all $\alpha_1, \ldots, \alpha_n$ such that $\Psi \vdash \alpha_k$ for all $k \in [1,n]$, if there exists an application of an inference rule $\iota \in \mathcal{I}_{\Psi,\gamma}$, such that $\frac{\alpha_1,\ldots,\alpha_n}{\beta}$ and the parameters of ι are specified by priorities $\pi_1, \ldots, \pi_m$ then

$$(\{(\pi_1 \wedge \cdots \wedge \pi_m) \rightarrow (\alpha_1 \wedge \cdots \wedge \alpha_n \rightarrow \beta)\} : (\pi_1 \wedge \cdots \wedge \pi_m) \rightarrow (\alpha_1 \wedge \cdots \wedge \alpha_n \rightarrow \beta))$$

is a bdu of $\mathcal{L}_{\text{Arg}}$.

In items 1 and 2 the relevant elements of the agent's reasoning are added to the argumentation language. In items 3 and 4 the tools for reasoning about trust are added: in 3 the trust priority rules and in 4 the rules of the trust model. The bdus added in 4 contain a double implication: they state that if an agent has the priorities in Π_L then a *trust rule* (which was a bdu in Pinyol's argumentative theory) holds. In practice what this accomplishes, is to allow the argumentation to go a level deeper: agents can now argue about *why* a trust rule, representing an application of a deduction rule in the trust model, holds. An argument for a trust evaluation can be represented in a tree. At each level, a node can be deduced by using the deduction rules of $\mathcal{L}_{\text{Arg}}$ with as preconditions the node's children. A leaf in the tree is a bdu. Each agent can construct its own argumentation tree for a trust evaluation and use this in a dialogue to communicate personalised trust evaluations.

3.2.2. *The dialogue protocol*

Koster, Sabater-Mir, and Schorlemmer (2012b) define a dialogue protocol, based largely upon a previous protocol for information-seeking dialogue (Prakken, 2005). The dialogue for personalising trust evaluations starts as an information-seeking dialogue, but if the agents discover their priorities are incompatible, they can discover whether this is due to a lack of information of either

agent, or whether their world views are simply incompatible. If either agent is lacking information or the agents think they can reach an agreement on beliefs, they can enter a persuasion dialogue to achieve an agreement on the beliefs and trust priority rules. If this succeeds, they can restart the dialogue and see if they now agree on trust evaluations. In this way the argument serves to allow cooperative agents to converge on a similar model of trust and supply each other with personalised trust recommendations.

The main issue with these systems is that they only work if they are adopted by a sufficient number of agents. In particular, the framework for personalised trust recommendations requires agents to use a cognitive agent model with the computational trust model incorporated into it, and is thus only applicable in domains where such a cognitive model is already necessary. However, in such domains, trust will necessarily be subjective, and some process, such as personalisation of trust recommendations, is necessary for communication about trust to be useful at all.

4. Discussion

This is the first survey on the combination of computational trust and argumentation. We identify the potential areas where trust can be combined with argumentation and discuss the existing approaches. Our aim was to serve as a guide for future development in the field, because while the fields of computational trust, and argumentation are themselves in rapid development, their combination seems like a natural match: both aim to deal with an uncertain, social environment, and their synergy is obvious. In Table 1 we give an overview of how the various approaches that we discussed use trust and argumentation.

One thing to note is that the division along the two separate axes works well for reasoning, but it is hard to see how communication would benefit from added trust in arguments. In particular, the models that deal with reasoning about trust in arguments can be seen as models for communication: the trustworthiness of an argument depends on the source of the information, and thus we could see this as a model for using trust in communication about arguments as well. However, the main focus of these works is on reasoning within a single agent.

A major concern is the computational aspect. In the table we have ticked the works that provide a clear computational model of the proposed approach. da Costa Pereira et al. and Tang et al. provide the algorithms necessary for implementation, but insofar as we know there is no implemented version. Even the ones that do discuss a prototype in a limited simulation environment. It is

Table 1. Summary of the works discussed, with regards to the principal axes of investigation, as well as whether they provide an implementation (T. and A. are short for Trust and Argumentation, respectively).

	Reasoning		Communication		
	T. in A.	A. about T.	T. in A.	A. about T.	Implementation provided
Koster et al. (2012a)				✓	✓
Matt et al. (2010)		✓			✓
Parsons et al. (2011)	✓				
da Costa Pereira et al. (2011)	✓				∼
Pinyol (2011)				✓	✓
Prade (2007)		✓			
Stranders et al. (2008)		✓		∼	✓
Tamargo et al. (2012)	✓				
Tang et al. (2012)	✓				∼
Villata et al. (2013)	✓	✓			

very much an open question how well they deal with the computational complexity of a real environment, in which they are designed to operate.

One aspect that has not received a lot of attention is the problem of communicating an agent's decision to the user. In particular when a decision must be made to trust one agent over another, the reasons must be clearly communicable to the user. Argumentation can help with this. The work by Stranders et al. (2008) is aimed specifically at this, but other works have visual representations of the arguments that could be seen as early prototypes on how to display this information to the user (Koster et al., 2012a; Tang et al., 2012; Villata et al., 2013). Nevertheless, while all these models allow for insight into the agent's reasoning process, a lot more work has to be done to make the communication to the user clear, or better yet, interactive.

Finally, the models for belief revision (da Costa Pereira et al., 2011; Tamargo et al., 2012; Tang et al., 2012) do not provide formal proofs about the logical properties of the revision operator they define. While they clearly reject some of the AGM postulates, it is not necessary to reject all of them, and some weaker version of the original postulates might be useful. In particular, such proofs could be used to guarantee that important properties of the agent's beliefs are not violated using these revision operators. Both da Costa Pereira et al. and Tamargo et al. prove some, but not all, desiderata of belief revision operators, whereas Tang et al. are more interested in the computational aspects. More work should be done in both investigating the formal properties of such operators, as well as finding desirable properties for belief revision operators that are better suited to a dynamic, uncertain environment.

Decision making in an uncertain, social environment is one of the hardest problems for computational systems to deal with, and the approaches discussed are not full-fledged solutions, but rather an exploration of possible ones. By describing the works and highlighting their strengths and weaknesses, we hope to point readers in fruitful directions for future research. In particular, creating *scalable* algorithms for generating trustworthy arguments is a problem and the principal hurdle in most of the work discussed.

Acknowledgements

Andrew Koster is supported by CAPES (PNPD). He wishes to thank Jordi Sabater-Mir and Marco Schorlemmer for their assistance in the early stages of this research, and Ana Bazzan for her support while writing this paper. Finally, a thank you to the anonymous reviewers for their feedback.

References

Alchourrón, C.E., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic, 50*, 510–530.

Amgoud, L., & Prade, H. (2004). Using arguments for making decisions: A possibilistic logic approach. In *Proceedings of the 20th conference on uncertainty in artificial intelligence (UAI'04)* (pp. 10–17). Banff: AUAI Press.

Barber, S.K., & Kim, J. (2001). Belief revision process based on trust: Agents evaluating reputation of information sources. In R. Falcone, M. Singh, & Y.H. Tan (Eds.), *Trust in cyber-societies: Vol. 2246. LNCS* (pp. 73–82). Barcelona: Springer.

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

Boella, G., Gabbay, D., van der Torre, L., & Villata, S. (2009). Meta-argumentation modelling I: Methodology and techniques. *Studia Logica, 93*, 297–355.

Caminada, M. (2006). On the issue of reinstatement in argumentation. In M. Fisher, W. van der Hoek, B. Konev, & A. Lisitsa (Eds.), *Proceedings of the 10th European conference on logics in artificial intelligence (JELIA'06): Vol. 4160. LNCS* (pp. 111–123). Liverpool: Springer.

Castelfranchi, C., & Falcone, R., (2010). *Trust theory: A socio-cognitive and computational model*. Chichester: Wiley.

Chesñevar, C., & Simari, G. (2007). Modelling inference in argumentation through labelled deduction: Formalization and logical properties. *Logica Universalis, 1*, 93–124.

da Costa Pereira, C., & Tettamanzi, A.G.B. (2010). An integrated possibilistic framework for goal generation in cognitive agents. In *Proceedings of the 9th international conference on autonomous agents and multiagent systems (AAMAS 2010)* (pp. 1239–1246). Toronto: IFAAMAS.

da Costa Pereira, C., Tettamanzi, A., & Villata, S. (2011). Changing one's mind: Erase or rewind? Possibilistic belief revision with fuzzy argumentation based on trust. In T. Walsh (Ed.), *Proceedings of IJCAI'11* (pp. 164–171). Barcelona: AAAI Press.

Dung, P.M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence, 7*, 321–358.

Falcone, R., & Castelfranchi, C. (2012). Trust and transitivity: How trust-transfer works. In *Proceedings of the 10th international conference on practical applications of agents and multi-agent systems (PAAMS'12): Vol. 156. Advances in intelligence and soft computing* (pp. 179–187). Salamanca: Springer.

Golbeck, J. (2006). Combining provenance with trust in social networks for semantic web content filtering. In L. Moreau & I. Foster (Eds.), *Provenance and annotation of data (IPAW 2006): Vol. 4145. LNCS* (pp. 101–108). Chicago, IL: Springer.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.

Hansson, S.O. (1999). A survey of non-prioritized belief revision. *Erkenntnis, 50*, 413–427.

Jones, N.D. (1997). *Computability and complexity: From a programming perspective*, Foundations of Computing Series. Cambridge, MA: MIT Press.

Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems, 43*, 618–644.

Koster, A., Sabater-Mir, J., & Schorlemmer, M. (2012a). Personalizing communication about trust. In *Eleventh international conference on autonomous agents and multiagent systems (AAMAS'12)* (pp. 517–526). Valencia, Spain: IFAAMAS.

Koster, A., Sabater-Mir, J., & Schorlemmer, M. (2012b). A formal argumentation dialogue for personalised trust communication. In *Proceedings of the fifteenth workshop 'trust in agent societies' at AAMAS '12* (pp. 55–66). Valencia, Spain: IFAAMAS.

Koster, A., Schorlemmer, M., & Sabater-Mir, J. (2013). Opening the black box of trust: Reasoning about trust models in a BDI agent. *Journal of Logic and Computation, 23*, 25–58.

Matt, P.A., Morge, M., & Toni, F. (2010). Combining statistics and arguments to compute trust. In *Proceedings of the 9th international conference on autonomous agents and multiagent systems (AAMAS 2010)* (pp. 209–216). Toronto: IFAAMAS.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*, 57–74.

Ossowski, S. (Ed.). (2013). *Agreement technologies: Vol. 8. Law, governance and technology*. Springer.

Paglieri, F., & Castelfranchi, C. (2006). The Toulmin test: Framing argumentation within belief revision theories. *Arguing on the Toulmin Model, 10*, 359–377.

Pardo, P., Pajares, S., Onaindia, E., Godo, L., & Dellunde, P. (2011). Multiagent argumentation for cooperative planning in DeLP-POP. In K. Tumer, P. Yolum, L. Sonenberg, & P. Stone (Eds.), *Proceedings of the 10th international conference on autonomous agents and multiagent systems (AAMAS 2011)* (pp. 971–978). Taipei: IFAAMAS.

Parsons, S., Tang, Y., Sklar, E., McBurney, P., & Cai, K. (2011). Argumentation-basded reasoning in agents with varying degrees of trust. In K. Tumer, P. Yolum, L. Sonenberg, & P. Stone (Eds.), *Proceedings of the 10th international conference on autonomous agents and multiagent systems (AAMAS 2011)* (pp. 879–886). Taipei: IFAAMAS.

Pinyol, I., (2011). *Milking the reputation cow: Argumentation, reasoning and cognitive agents: Vol. 44. Monografies de l'Institut d'Investigació en Intel·ligencia Artificial*. Consell Superior d'Investigacions Científiques.

Pinyol, I., & Sabater-Mir, J. (2009). Towards the definition of an argumentation framework using reputation information. In *Proc. of the Twelfth Workshop 'Trust in Agent Societies' at AAMAS '09* (pp. 92–103). Budapest: IFAAMAS.

Pinyol, I., & Sabater-Mir, J. (2013). Computational trust and reputation models for open multi-agent systems: A review. *Artificial Intelligence Review, 30*, 1–25.

Pollock, J.L. (1987). Defeasible reasoning. *Cognitive Science, 11*, 481–518.

Prade, H. (2007). A qualitative bipolar argumentative view of trust. In V. Subrahmnanian & H. Prade (Eds.), *International conference on scalable uncertainty management (SUM 2007): Vol. 4772. LNAI* (pp. 268–276). Washington, DC: Springer.

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

Prakken, H. (2006). Formal systems for persuasion dialogue. *Knowledge Engineering Review, 21*, 163–188.

Rahwan, I. (2006). Guest editorial: Argumentation in multi-agent systems. *Autonomous Agents and Multi-Agent Systems, 11*, 115–125.

Rao, A.S., & Georgeff, M.P. (1991). Modeling rational agents within a BDI-architecture. In R. Fikes & E. Sandewall (Eds.), *Proceedings of the 2nd international conference on principles of knowledge representation and reasoning* (pp. 473–484). San Mateo, CA: Morgan Kaufmann.

Stranders, R., de Weerdt, M., & Witteveen, C. (2008). Fuzzy argumentation for trust. In F. Sadri & K. Satoh (Eds.), *Computational logic in multi-agent systems: 8th international workshop, CLIMA VIII: Vol. 5056. LNCS* (pp. 214–230). Porto: Springer.

Tamargo, L.H., García, A.J., Thimm, M., & Krümpelmann, P. (2012). Selective revision with multiple informants and argumentative support. *Revista Iberoamericana de Inteligencia Artificial, 15*, 4–17.

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

Villata, S., Boella, G., Gabbay, D.M., & van der Torre, L. (2013). A socio-cognitive model of trust using argumentation theory. *International Journal of Approximate Reasoning, 54*, 541–559.

Walton, D.N. (1990). What is reasoning? What is an argument? *Journal of Philosophy, 87*, 399–419.

Walton, D.N., & Krabbe, E.C.W. (1995). *Commitment in dialogue, basic concepts of interpersonal reasoning*. Albany, NY: SUNY Press.

Yu, B., & Singh, M.P. (2002). Distributed reputation management for electronic commerce. *Computational Intelligence, 18*, 535–549.