Trusting the Messenger and the Message

S. Villata[1], F. Paglieri[2], A. Tettamanzi[3],
R. Falcone[2], C. da Costa Pereira[4], and C. Castelfranchi[2]

[1] INRIA Sophia Antipolis, serena.villata@inria.fr
[2] ISTC CNR Roma, fabio.paglieri@istc.cnr.it,
rino.falcone@istc.cnr.it, cristiano.castelfranchi@istc.cnr.it
[3] Università degli Studi di Milano, Dipartimento di Tecnologie dell'Informazione
andrea.tettamanzi@unimi.it
[4] Université de Nice Sophia Antipolis, I3S Lab (UMR CNRS 7271)
celia.pereira@unice.fr

Abstract. Information provided by a source should be assessed by an intelligent agent on the basis of several criteria: most notably, its content and the trust one has in its source. In turn, the observed quality of information should feed back on the assessment of its source, and such feedback should intelligently distribute among different features of the source—e.g., competence and sincerity. We propose a formal framework in which trust is not treated as a monolithic and static concept. We regard trust as a multi-dimensional concept relativized to the sincerity of the source and its competence with respect to specific domains: both these aspects influence the assessment of the information, and also determine a feedback on the trustworthiness degree of its source. We provide a framework to describe the combined effects of competence and sincerity on the perceived quality of information. We focus on the feedback dynamics from information quality to source evaluation, highlighting the role that uncertainty reduction, and social comparison play in determining the amount and the distribution of feedback.

Category: I.2.11, Distributed Artificial Intelligence, Intelligent agents

Terms: Theory

Keywords: Knowledge representation, Cognition, Reasoning

1 Introduction

In real life, the trust assigned to a message depends crucially, albeit not solely, on the perceived trustworthiness of its source [7, 8]. In turn, whether or not the message turns out to be reliable has important repercussions on the source trustworthiness. This is also true with respect to the exchange of arguments in social interaction. When people argue with each other, trying to get their arguments

accepted to foster their own goals, they also evaluate the arguments proposed by other parties in the discussion. This evaluation considers not only the possible conflicts with other arguments, but also the trustworthiness degree of the information source proposing the arguments. In argumentation theory [15], the arguments are considered to be accepted or not depending on the attacks against them. In this kind of frameworks, neither the information sources proposing the arguments nor their trustworthiness degree are considered. In recent years, the area has seen a number of proposals [14, 16, 11, 13, 18, 3] to introduce the trust component in the evaluation process of the arguments. The common drawback of these approaches is that they do not return the intrinsic complexity of the trust notion, as highlighted instead by socio-cognitive models.

In this paper, we adopt the socio-cognitive model of trust proposed by Castelfranchi and Falcone [2], and we elaborate its computational counterpart, although with some expressivity limitations due to the overall complexity of the model. We are interested in investigating how that socio-cognitive model of trust can be extended to a model based on argumentation. This breaks down into the following sub-questions:

- How to distinguish different dimensions of trust, e.g., sincerity and competence, and model their respective contribution?
- How to model the trust feedback from arguments to information sources?

We address these questions starting from the argumentation-based model for belief revision recently proposed by da Costa Pereira et al. [3], which originally lacked a representation of a cognitive model of trust and its inherent dynamics. We extend that model along the following lines.

First, trust is not a monolithic concept. We relativize the notion of trust to the dimensions of sincerity and competence in various domains. For instance, a reliable motor mechanic will be considered competent in the cars domain, but not necessarily so when suggesting the best restaurant to eat pizza; conversely, a pizza maker is typically assumed to be trustworthy on the latter domain but not on the former. The trust in the information source is not absolute, but it is relative to an estimate of sincerity and competence in the relevant domain. Sincerity and competence thus combine to produce an aggregated degree of trust.

Second, trust is not a static concept. There is a bidirectional link between the source and its information items. This means that an argument is more or less believable on the basis of the source's trustworthiness, but this leads to a feedback such that the invalidation of the argument, due to attacks by other trustworthy arguments, feeds back on the source's credibility. The sign of the feedback depends on how much the "quality" of the message surprises the agent w.r.t. its prior assessment of the source trustworthiness.

The paper is organized as follows: Section 2 highlights the main differences of our approach with related work. Section 3 provides the basic concepts of the model proposed by Pereira et al. [3]. Sections 4 and 5 introduce the multidimensional trust model and specify the feedback mechanism. Conclusions end the paper.

2 Related Work

The importance of relating trust and argumentation has been underlined by Dix et al. [4], who present trust as a major issue concerning the research challenges for argumentation. Also Parsons et al. [12] present the provenance of trust as one of the mechanisms to be investigated in argumentation. They claim that a problem, particularly of abstract approaches such as Dung [6], is that they cannot express the provenance of trust, and the fact that b is attacked because b is proposed by agent s who is not trustworthy. Starting from this observation, we propose a model of argumentation where the arguments are related to the sources and their degree of acceptability is computed on the basis of the trustworthiness degree of the sources. Furthermore, our approach goes beyond this observation by providing a feedback such that the final quality of the arguments influences the source evaluation as well.

Most studies in this domain used argumentation to model trust dynamics and/or reasoning about trust (e.g., [14, 11, 16, 18]), which is a worthy but different enterprise from the one pursued here. Closer in spirit to the present paper is the work by Parsons et al. [17, 13], who present a framework to introduce the sources in argumentation and to express the degrees of trust. They define trust-extended argumentation graphs in which each premise, inference rule, and conclusion is associated to the trustworthiness degree of the source proposing it. Thus, given two arguments rebutting each other, the argument whose conclusion has a higher trust value is accepted. The difference is that in such a framework the trust values associated to the arguments do not change and the arguments are accepted with the same degree even if they are attacked by more trusted arguments. Again, the feedback towards the source as well as the distinction between competence and sincerity is not considered.

A huge amount of research has been conducted on trust, and some of these works are described below, even if in this paper we limit our attention to the cognitive trust model of Castelfranchi and Falcone [2]. An approach to model trust using modal logic is proposed by Lorini and Demolombe [10], who present a concept of trust that integrates the trusters goal, the trustees action ensuring the achievement of the trusters goal, and the trustees ability and intention to do this action—taking again inspiration from [2]. Another proposal is presented by Liau [9], in which the influence of trust on the assimilation of information into the source's mind is considered. The idea is that "if agent i believes that agent j has told him the truth on p, and he trusts the judgement of j on p, then he will also believe p". Wang and Singh [20], instead, understand trust in terms of belief and certainty: A's trust in B is reflected in the strength of A's belief that B is trustworthy. They formulate certainty in terms of evidence based on a statistical measure defined over a probability distribution of positive outcomes. Both Liau [9] and Wang and Singh [20] capture intuitions that play a role also in our approach, but they vastly oversimplify the nature and dynamics of trust, as opposed to the socio-cognitive model discussed in Castelfranchi and Falcone [2].

3 Background

A classical propositional language may be used to represent information for manipulation by a cognitive agent.

Definition 1. *(Language) Let* Prop *be a* finite*. set of atomic propositions and let* $\mathcal{L}$ *be the propositional language such that* $\text{Prop} \cup \{\top, \bot\} \subseteq \mathcal{L}$*, and,* $\forall \phi, \psi \in \mathcal{L}$*,* $\neg\phi \in \mathcal{L}$*,* $\phi \wedge \psi \in \mathcal{L}$*,* $\phi \vee \psi \in \mathcal{L}$*.*

As usual, one may define additional logical connectives and consider them as useful shorthands for combinations of connectives of $\mathcal{L}$, e.g., $\phi \supset \psi \equiv \neg\phi \vee \psi$. We denote by $\Omega = \{0,1\}^{\text{Prop}}$ the set of all interpretations on Prop. An interpretation $\mathcal{I} \in \Omega$ is a function $\mathcal{I} : \text{Prop} \to \{0,1\}$ assigning a truth value $p^{\mathcal{I}}$ to every atomic proposition $p \in \text{Prop}$ and, by extension, a truth value $\phi^{\mathcal{I}}$ to all formulas $\phi \in \mathcal{L}$. We denote by $[\phi]$ the set of all models of ϕ, $[\phi] = \{\mathcal{I} : \mathcal{I} \models \phi\}$.

A Dung's abstract argumentation framework [6] (AF) is a pair $\langle \mathcal{A}, \rightarrow \rangle$ where $\mathcal{A}$ is a set of elements called *arguments* and $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is a binary relation called *attack*. Dung defines a number of acceptability semantics [6] to assess which are the sets of accepted arguments. We can give arguments a structure, and the attack relation is defined in terms of such a structure of the arguments: an argument is a pair $A = \langle \Phi, \phi \rangle$, with $\phi \in \mathcal{L}$ and $\Phi \subset \mathcal{L}$, such that (i) $\Phi \nvdash \bot$, (ii) $\Phi \vdash \phi$, (iii) Φ is minimal w.r.t. set inclusion. We call ϕ the conclusion and Φ the support of the argument. Given $A \in \mathcal{A}$, $\text{src}(A)$ is the set of the sources of A.

In a recent paper, da Costa Pereira et al. [3] propose a framework where argumentation theory is used in belief revision. In this framework, the arguments are weighted on the basis of the trustworthiness degree of the agents proposing them. The acceptability of the arguments is then computed by a labelling algorithm which assigns the arguments a fuzzy value, differently from Dung-like frameworks where arguments are either accepted or rejected. We select this work as the basis of our trust model because it provides (i) an explicit link between the trustworthiness degree of the sources and of the arguments, (ii) a mechanism such that the initial value assigned to the arguments changes due to the attacks against them, (iii) the beliefs of the agents are also involved, not only their arguments.

Given an AF and the trust degree τ_s of each source s, the labeling algorithm [3] computes a fuzzy extension as a fuzzy set of accepted arguments, whose membership α assigns to each argument A a degree of acceptability $\alpha(A)$ such that $\alpha(A) = 0$ means the argument is outright unacceptable, $\alpha(A) = 1$ means the argument is fully acceptable, and all cases inbetween are provided for. Then the labeling α is used to determine the agent's beliefs, by constructing a possibility distribution from which the degree of belief of an arbitrary formula may be calculated.

A possibility distribution may be defined as the membership function of a fuzzy set that describes the more or less possible and mutually exclusive values of one (or more) variable(s) [21]. Indeed, if F designates the fuzzy set of possible values of a variable X, $\pi_X = \mu_F$ is called the possibility distribution associated

to X. The identity $\mu_F(v) = \pi_X(v)$ means that the membership degree of v to F is equal to the possibility degree of X being equal to v when all we know about X is that its value is in F. A possibility distribution for which there exists a completely possible value ($\exists v_0 : \pi(v_0) = 1$) is said to be *normalized*.

Definition 2. *(Possibility and Necessity Measures) A possibility distribution π induces a* possibility measure *and its dual* necessity measure*, denoted by Π and N respectively. Both measures apply to a crisp set A and are defined as follows:*

$$\Pi(A) = \sup_{s \in A} \pi(s); \tag{1}$$

$$N(A) = 1 - \Pi(\bar{A}) = \inf_{s \in \bar{A}} \{1 - \pi(s)\}. \tag{2}$$

In words, the possibility measure of A corresponds to the greatest of the possibilities associated to its elements; conversely, the necessity measure of A is equivalent to the impossibility of its complement $\bar{A}$.

The beliefs of an agent are thus completely described by a normalized possibility distribution $\pi : \Omega \to [0, 1]$, which represents a plausibility order of possible states of affairs: $\pi(\mathcal{I})$ is the possibility degree of interpretation $\mathcal{I}$.

Given $A = \langle \Phi, \phi \rangle$, let $\mathrm{con}(A)$ denote the conclusion of A, i.e., $\mathrm{con}(\langle \Phi, \phi \rangle) = \phi$. The possibility distribution π induced by a fuzzy labeling α is constructed by letting, for all interpretation $\mathcal{I}$,

$$\pi(\mathcal{I}) = \min\{1, 1 + \max_{A:\mathcal{I} \models \mathrm{con}(A)} \alpha(A) - \max_{B:\mathcal{I} \not\models \mathrm{con}(B)} \alpha(B)\}, \tag{3}$$

where the first maximum accounts for the most convincing argument compatible with $\mathcal{I}$, and the second maximum accounts for the most convincing argument against $\mathcal{I}$. A world is possible to an extent proportional to the difference between the most convincing argument supporting it and the most convincing argument against it. The world is considered completely possible if such difference is positive or null, but it is considered less and less possible (or plausible) as such difference grows more and more negative.

The degree to which a given arbitrary formula $\phi \in \mathcal{L}$ is believed is calculated from the possibility distribution induced by the fuzzy argumentation framework as $\mathbf{B}(\phi) = N([\phi]) = 1 - \max_{\mathcal{I} \not\models \phi}\{\pi(\mathcal{I})\}$, where $\mathbf{B}$ may be regarded, at the same time, as a fuzzy modal epistemic operator or as a fuzzy subset of $\mathcal{L}$. Notice that $\mathbf{B}(\phi)$ can be computed for any formula ϕ, not just for formulas that are the conclusion of some argument. E.g., if A is an argument whose conclusion is p and B is an argument whose conclusion is $p \supset q$, and $\alpha(A) = \alpha(B) = 1$, then not only $\mathbf{B}(p) = \mathbf{B}(p \supset q) = 1$, but also $\mathbf{B}(q) = 1$, $\mathbf{B}(p \wedge q) = 1$, etc. Consequences of the properties of possibility and necessity measures are that $\mathbf{B}(\phi) > 0 \Rightarrow \mathbf{B}(\neg\phi) = 0$, which means that if the agent somehow believes ϕ then it cannot believe $\neg\phi$ at all, and

$$\mathbf{B}(\top) = 1, \tag{4}$$

$$\mathbf{B}(\bot) = 0, \tag{5}$$

$$\mathbf{B}(\phi \wedge \psi) = \min\{\mathbf{B}(\phi), \mathbf{B}(\psi)\}, \tag{6}$$

$$\mathbf{B}(\phi \vee \psi) \geq \max\{\mathbf{B}(\phi), \mathbf{B}(\psi)\}. \tag{7}$$

4 Trust model

In this section, we take a look into the multidimensional nature of trust. For the sake of brevity, we simplify Castelfranchi and Falcone's model [2], and focus only on two broad categories of relevant features in the source: competence (to what extent the source is deemed able to deliver a correct argument) and sincerity (to what extent the source is considered willing to provide a correct argument), both of which contribute to determine the source's overall trustworthiness. Importantly, evaluations of competence and sincerity are allowed to change across different domains. For instance, I might think that the colleague who is competing with me for a promotion is likely to be insincere in giving me tips on how to improve my career, and yet there is no reason to doubt his sincerity when he suggests me what movie to watch tonight.

Here, we consider competence and sincerity as two possible dimensions for measuring the trustworthiness of a source. Such dimensions will be represented as graded beliefs, that is, an agent believes, to a given degree, that a source is competent (sincere) with respect to a domain. In order to represent the fact that the agent's beliefs can be incomplete (an agent may not have beliefs on everything), it is important to make it clear that one may either "believe p", "believe $\neg p$", or believe none of them, due to ignorance [5]. In such a case, it must be possible to make the fact that $\mathbf{B}(p) = 0$ and $\mathbf{B}(\neg p) = 0$ explicit. This is the reason why the competence and sincerity will be represented by a bipolar pair of beliefs, in competence (sincerity) and in the negation thereof.

The idea is that each source is assigned by the agent a belief of competence and another belief of sincerity, both represented by bipolar values ranging between 0 and 1. These beliefs combine to determine the degree of trustworthiness. What is crucial is that the basic properties of sincerity and competence are kept separate, since the feedback will be designed to affect differently each of them, and only as a consequence will also impact on the source's aggregate trustworthiness.

4.1 Modeling Competence

Here, the idea is to define a number of competence domains, with respect to which (i) the competence of a source s is evaluated, yielding the trust an agent has in s when it offers arguments relevant to any domain—the trust in the competence of s is a vector $\mathbf{c}(s)$ whose elements may be formally regarded as bipolar degrees of belief $\langle c_d^+, c_d^- \rangle$, where c_d^+ is the degree to which the agent believes the source *is* competent about domain d, and c_d^- is the degree to which the agent believes the source *is not* competent about d[1], and (ii) the positive and negative feedback reflected on the information source is also domain-specific, e.g., if the doctor tells the agent to go to a restaurant that turns out to be bad,

[1] c_d^+ and c_d^- obey the property, typical of necessities and beliefs, that $c_d^+ > 0 \Rightarrow c_d^- = 0$ and, *vice versa*, $c_d^- > 0 \Rightarrow c_d^+ = 0$.

this reduces the agent's confidence in her gastronomic taste, not in her medical skills.

We propose to associate the arguments, by way of the formula in their conclusion, to the competence domains depending on the topic their are talking about. The atomic propositions of the language are mapped to the competence domains with a certain degree, i.e., the degree to which the atomic proposition belongs in the specific domain. The degree to which a propositional formula (argument) belongs in a specific domain might be given by the maximum degree to which the atomic propositions on which the truth value of the formula depends belong to the competence domain.

Definition 3. *Given D the set of competence domains, and* Prop *the set of atomic propositions, the association between atoms and domains, is represented by a fuzzy relation $R \subseteq D \times \text{Prop}$. Given $p \in \text{Prop}$, $d \in D$, the membership degree of the atomic proposition p to domain d is $R(d, p) \in [0, 1]$.*

We now extend R of Definition 3 to $D \times \mathcal{L}$.

Definition 4. *Let $\phi \in \mathcal{L}$ be a propositional formula and $d \in D$ be a competence domain, then*

$$R(d, \phi) = \max_{p \in \text{DS}(\phi)} R(d, p),$$

where $\text{DS}(\phi) = \{p \in \text{Prop} : \exists \mathcal{I} \models p, \mathcal{I}' \not\models p : \phi^{\mathcal{I}} \neq \phi^{\mathcal{I}'}\}$ *is the* determinant set *of ϕ, i.e., the set of all atomic propositions on which the truth of ϕ depends.*

4.2 Modeling Sincerity

The notion of *sincerity* is a property typically attributed to agents with goals and intentions. Talking about the sincerity of an information source is assuming that source is another agent and has intentions, which can be in harmony or in conflict with the goals of the recipient. For example, when I meet my bank's personal investment advisor to get advice on possible placements of my savings, I should know from the outset that, among her goals, she has the one of maximizing the profits of her employer. Therefore, I may expect that she will be tempted to manipulate my beliefs to lure me into buying financial instruments on which the bank makes a profit.

Furthermore, in general, the sincerity of a source should relate differently to each individual domain: a malicious agent may have an advantage to lie about a domain somehow related with its goals, but has no interest in lying about unrelated domains. Therefore, it would be wrong to regard sincerity as an absolute property of a source. Thus, both competence and sincerity are relativized to individual domains.

Based on the above discussion, we will model the beliefs an agent maintains about the sincerity of its sources as a vector $\sigma(s)$, whose components, associated to individual domains, are bipolar values $\langle \sigma_d^+, \sigma_d^- \rangle$ where $\sigma_d^+ \in [0, 1]$ represents the degree to which the agent has reasons to believe that source s is sincere about

d, whereas $\sigma_d^- \in [0,1]$ represents the degree to which the agent has reasons to believe the contrary; $\sigma_d^+ = \sigma_d^-$ represents a status of maximal uncertainty about the sincerity of s about d. Since σ_d^+ and σ_d^- represent beliefs, they must satisfy all properties of beliefs, in particular that $\sigma_d^+ > 0 \Rightarrow \sigma_d^- = 0$ and $\sigma_d^- > 0 \Rightarrow \sigma_d^+ = 0$.

4.3 Aggregating Competence and Sincerity

Competence and sincerity may be aggregated into a single trust value τ_d, used to weight arguments, by recalling that both concepts are formally beliefs, although of a special kind, not induced by the agent's AF, but determined by the internal mechanisms of competence and sincerity evaluation.

In particular, one might argue that a source is trusted to the extent that it is believed to be both competent and sincere. If this is the intended meaning of trust, then, for all domains d, the degree to which s is trusted about d is given by

$$\begin{aligned}\tau_d(s) &= \mathbf{B}(\text{competent}(d,s) \wedge \text{sincere}(d,s)) \\ &= \min(\mathbf{B}(\text{competent}(d,s)), \mathbf{B}(\text{sincere}(d,s))),\end{aligned}$$

where competent(d,s) means that s is competent about d and sincere(d,s) means that s is sincere about d.

This belief degree may be computed by reconstructing the possibility distribution π that induces the beliefs

$$\begin{aligned}\mathbf{B}(\text{competent}(d,s)) &= c_d^+, \\ \mathbf{B}(\neg\text{competent}(d,s)) &= c_d^-, \\ \mathbf{B}(\text{sincere}(d,s)) &= \sigma_d^+, \\ \mathbf{B}(\neg\text{sincere}(d,s)) &= \sigma_d^-.\end{aligned}$$

There are four possible worlds, as far as the competence and sincerity of source s about d goes, namely

$$\begin{aligned}\mathcal{I}_0 &= \text{competent}(d,s)) \wedge \text{sincere}(d,s)), \\ \mathcal{I}_1 &= \text{competent}(d,s)) \wedge \neg\text{sincere}(d,s)), \\ \mathcal{I}_2 &= \neg\text{competent}(d,s)) \wedge \text{sincere}(d,s)), \\ \mathcal{I}_3 &= \neg\text{competent}(d,s)) \wedge \neg\text{sincere}(d,s)).\end{aligned}$$

Let us abbreviate $\pi(\mathcal{I}_i)$ as π_i. Since we know that

$$\begin{aligned}\max\{\pi_0, \pi_1\} &= 1 - c_d^-, \\ \max\{\pi_2, \pi_3\} &= 1 - c_d^+, \\ \max\{\pi_0, \pi_2\} &= 1 - \sigma_d^-, \\ \max\{\pi_1, \pi_3\} &= 1 - \sigma_d^+,\end{aligned}$$

we may solve this system of four equations for the four unknown variables $\pi_0, \pi_1, \pi_2, \pi_4$ to get

$$\begin{aligned}
\pi_0 &= 1 - \max\{c_d^-, \sigma_d^-\}, \\
\pi_1 &= 1 - \max\{c_d^-, \sigma_d^+\}, \\
\pi_2 &= 1 - \max\{c_d^+, \sigma_d^-\}, \\
\pi_3 &= 1 - \max\{c_d^+, \sigma_d^+\}.
\end{aligned}$$

Therefore, the single trust value for domain d is given by the conjunction of how much the agent believes source s is competent and sincere concerning d

$$\begin{aligned}
\tau_d(s) &= \mathbf{B}(\text{competent}(d,s) \wedge \text{sincere}(d,s)) = \\
&= 1 - \Pi([\neg\text{competent}(d,s) \vee \neg\text{sincere}(d,s)]) = \\
&= 1 - \max\{\pi_1, \pi_2, \pi_3\} = \\
&= \min\{\max\{c_d^-, \sigma_d^+\}, \max\{c_d^+, \sigma_d^-\}, \max\{c_d^+, \sigma_d^+\}\}.
\end{aligned}$$

5 Feedback Dynamics

In this section, we define the feedback dynamics on the sources. As we discussed in the previous sections, the acceptability of the arguments in our model depends on (i) the trustworthiness of the information source proposing the arguments, and (ii) the interactions of the proposed arguments and the other beliefs of the agent. What we want to introduce in this section is the idea that the final acceptability value of the arguments provides a feedback on the trustworthiness degree in the information source from the next interaction.

5.1 Overall Feedback: The role of prediction and surprise

The overall amount and sign (increment or decrement) of the feedback depends on how much the overall quality of the message surprises the agent, with respect to its prior assessment of the source trustworthiness. This captures the principle that information quality should change one's assessment of its source *only when the agent learns something new about the capacity* of the source to deliver information of either high or low quality. In other words, there should be a feedback on the source only when the quality of its argument tells me something new about the source's trustworthiness, revealing my previous opinion to be wrong. Otherwise, the quality of the new argument just *confirms* my previous assessment of the source, and confirmation, by definition, consolidates a pre-existing judgment, rather than modifying it. This points to the role of *prediction* in feedback dynamics from arguments to sources, and this prediction is based on the pre-existing degree of trustworthiness of the source of a given argument.

Let $\tau(s)$ be the current degree of trustworthiness of source s and $\mathbf{B}(\text{con}(A))$ the degree of belief, in light of current evidence, in argument A provided by s[2].

[2] $\text{con}(A)$ represent the conclusion of argument A.

Assuming that argument quality Q(A) is given by $\mathbf{B}(\mathrm{con}(A))$, then the total amount of feedback F_A produced by argument A on source s is given by

$$F_A = Q(A) - \tau(s). \tag{8}$$

The overall feedback F_A in our framework ranges between -1 (utter disappointment) and $+1$ (wonderful surprise), and goes to 0 whenever source s provides an argument A whose quality is exactly as expected ($Q(A) = \tau(s)$).

The critical point is that this feedback might affect to a different extent the two components of trustworthiness, to wit, competence and sincerity, depending on the agent's interpretation of what determined it. Feedback on sources is often specific: the agent does not only register the fact that the source provided information of good (or bad) quality, but it also *diagnoses* what virtue (or vice) prompted the source to do so. For instance, when I find out that a trusted source provided a poor piece of advice, should I conclude that the source was being deliberately insincere, or should I attribute the incident to a lack of competence? The overall trustworthiness is lowered in either case, but for very different reasons. Conversely, when a poorly estimated source provides surprisingly good information, is this because it stopped deceiving me, or because it became more competent?

5.2 Feedback Distribution: Uncertainty reduction and social comparison

Feedback dynamics face the problem of how to distribute the overall feedback F_A between competence and sincerity. In this respect, our current formal framework faces an important limitation: there is no way to access semantically the source's beliefs and goals, so they cannot be invoked to justify the different diagnoses the agent could make of information quality—contrary to what happens in real life, as discussed in [2]. For example, if I attribute to source s a goal which is currently in conflict with my own, any poor quality information I receive from s is more likely to be imputed to dishonesty than incompetence. In view of this limitation, the best we can do is to use smart rules-of-thumb that capture interesting (albeit partial) feedback regularities, leaving to future work further refinements on this point.

We propose two independent principles, whose effects sum up in distributing the feedback between competence and sincerity. The first principle focuses on the individual's previous assessment of a specific source, whereas the second principle compares what the single source is saying to what other sources said, and uses this degree of convergence to shape the feedback distribution. The first principle might be labelled *uncertainty reduction*: the idea is that a feedback should affect more the dimension for which there is greater uncertainty, that is, such that $U_c = 1 - \max\{c^+, c^-\}$, for competence, or $U_\sigma = 1 - \max\{\sigma^+, \sigma^-\}$, for sincerity, is greater. This works well for cases where the agent has formed a clear judgment on one dimension but is in doubt on the other: if I am convinced of your honesty but do not know whether you are competent or not (or *vice*

versa), then the quality of your argument is likely to be interpreted as evidence for or against your competence (or honesty)—after all, that is what I was not sure about to start with.

Given the total amount of feedback F_A from argument A, let $c = \langle c^+, c^- \rangle$ be the prior beliefs in competence, $\sigma = \langle \sigma^+, \sigma^- \rangle$ be the prior beliefs in sincerity, and $F_A(c)$ and $F_A(\sigma)$ the amount of feedback assigned to, respectively, c and σ. Then the following rule is used to capture the principle of uncertainty reduction:

$$F_A(c) = \frac{F_A U_c}{U_c + U_\sigma}, \quad F_A(\sigma) = \frac{F_A U_\sigma}{U_c + U_\sigma}. \tag{9}$$

This formulation satisfies the following desirable properties: (i) $F_A(c) + F_A(\sigma) = F_A$, (ii) $U_c > U_\sigma \Rightarrow |F_A(c)| > |F_A(\sigma)|$, (iii) $U_c < U_\sigma \Rightarrow |F_A(c)| < |F_A(\sigma)|$, and (iv) $U_c = U_\sigma \Rightarrow F_A(c) = F_A(\sigma) = \frac{1}{2} F_A$.

The second principle might be labeled *social comparison*: the idea is to compute the degree of *convergence* of an argument, measured as the number of sources that present that argument or arguments that support it, minus all the sources that are proposing arguments in contrast with it. Then convergence and quality of argument A are compared to determine what dimension of trustworthiness is more affected by the feedback, as follows:

(i) if A is *good* and *convergent*, there is a *stronger positive feedback on sincerity* than on competence (each source vouches for the sincerity of the other, even if they should all turn out to be mistaken);
(ii) if A is *good* and *divergent*, there is a *stronger positive feedback on competence* than on sincerity (an isolated source going against popular wisdom is likely to be on to something, like the biblical *vox clamantis in deserto*);
(iii) if A is *poor* and *convergent*, there is a *stronger negative feedback on competence* than on sincerity (it is unlikely that everybody is conspiring to fool you, whereas it is more plausible that they are all honestly mistaken);
(iv) if A is *poor* and *divergent*, there is a *stronger negative feedback on sincerity* than on competence (this is the typical case of a malicious or derailed source)[3].

More precisely, let Pro_A be the number of sources claiming argument A or supporting it and Con_A be the number of sources attacking argument A. Then we measure the degree of convergence k_A for A as follows:

$$k_A = \frac{\mathrm{Pro}_A - \mathrm{Con}_A}{\mathrm{Pro}_A + \mathrm{Con}_A}. \tag{10}$$

[3] Note that these principles are based on a number of *assumptions* (most notably, high level of independence and low probability of collusion among sources), and thus are not meant to be universally valid. Rather, they exemplify how simple rules-of-thumb can be identified to regulate feedback distribution, even without any explicit representation of context or agent's mental states. Testing their validity across various communicative situations (e.g., how much collusion is required to make these heuristics ineffective?) is left as future work.

In the limiting case, when there is no source either supporting or attacking A, this indicates that the argument has no external source, hence convergence does not apply. Otherwise, k_A always ranges between 1 (only supporting sources) and -1 (only attacking sources), with the 0 value indicating instances where the same number of sources support and attack A. Then the following rule is used to capture the principle of social comparison:

- if $k_A > 0$ and $F_A > 0$ (convergent argument producing a positive feedback), then the positive product $k_A F_A(c)$ is added to $F_A(\sigma)$ and subtracted from $F_A(c)$.
- if $k_A < 0$ and $F_A > 0$ (divergent argument producing a positive feedback), then the positive product $k_A F_A(\sigma)$ is added to $F_A(c)$ and subtracted from $F_A(\sigma)$.
- if $k_A > 0$ and $F_A < 0$ (convergent argument producing a negative feedback), then the negative product $k_A F_A(\sigma)$ is added to $F_A(c)$ and subtracted from $F_A(\sigma)$.
- if $k_A < 0$ and $F_A < 0$ (divergent argument producing a negative feedback), then the negative product $k_A F_A(c)$ is added to $F_A(\sigma)$ and subtracted from $F_A(c)$.
- finally, if $k_A = 0$ (neither convergent nor divergent), then social comparison has no effect on feedback distribution.

It is worth noting that the combination of both principles determines a distribution of the feedback between competence and sincerity that respects the following constraints:

$$0 \leq |F_A(c)| \leq |F_A| \quad \text{and} \quad 0 \leq |F_A(\sigma)| \leq |F_A|. \tag{11}$$

This states that distributing the feedback between competence and sincerity neither changes the overall amount of feedback, nor modifies its sign. At most, all the feedback will apply to only one dimension and not at all to the other: this happens, for instance, when the argument is either completely convergent ($k_A = 1$) or divergent ($k_A = -1$), but it never happens that the sum of the feedback on sincerity and the feedback on competence exceeds the overall amount of feedback produced by the argument, nor that the same argument generates a positive feedback on one dimension and a negative feedback for the other one. This is consistent with basic intuitions on how feedback ought to happen.

All in all, the system of rules described in this section provides a relatively simple way to characterize feedback dynamics from information quality to source evaluation, allowing to discriminate multiple dimensions of trustworthiness and capturing some intuitions on how feedback should be distributed among them. We certainly do not claim that these rules-of-thumb are perfect or immune to counter-examples—quite the contrary. However, they strike us as a useful and productive approximation of regularities in feedback dynamics, given current formal limitations in providing a semantic link between context of interaction, agents mental states (beliefs and goals) and argument assessment.

5.3 Feedback Application

We now have all the elements required to define how the feedback produced by argument A on source s is applied to the vectors $\mathbf{c}(s)$ and $\sigma(s)$.

For the sake of generality, let us denote by $\langle x^+, x^- \rangle$ the bipolar degrees that must be updated, by f the relevant dimension of feedback, and by $\langle y^+, y^- \rangle$ the updated degrees. One may regard $\langle x^+, x^- \rangle$ as formally equivalent to a variable $x \in [-1, 1]$, defined as $x = x^+ - x^-$, such that

$$x^+ = \begin{cases} x, \text{ if } x > 0, \\ 0 \quad \text{otherwise;} \end{cases} \tag{12}$$

$$x^- = \begin{cases} -x, \text{ if } x < 0, \\ 0 \quad\quad \text{otherwise.} \end{cases} \tag{13}$$

Now, the problem of updating the bipolar degrees of belief $\langle c_d^+, c_d^- \rangle$ with the competence dimension of feedback $F_A(c)$ and $\langle \sigma_d^+, \sigma_d^- \rangle$ with the sincerity dimension of feedback $F_A(\sigma)$ may be approached by mapping the bipolar degrees to the single variable $x = x^+ - x^-$ and compute

$$y = \begin{cases} x + (1 - x) \cdot f, \text{ if } f \geq 0, \\ x + (1 + x) \cdot f, \text{ if } f < 0, \end{cases} \tag{14}$$

from which $\langle y^+, y^- \rangle$ may be obtained by means of Equations 12 and 13. Feedback on competence and sincerity of s about d will be obtained by

$$F_A^d(c) = F_A(c) \cdot R(d, \text{con}(A)), \tag{15}$$

$$F_A^d(\sigma) = F_A(\sigma) \cdot R(d, \text{con}(A)). \tag{16}$$

Equations 15 and 16 amount to a *projection* of either dimension of the feedback along the domains related to the proposed argument, so that only the competence and sincerity about those domains are affected by the feedback.

6 Conclusions

Building on the socio-cognitive model of trust described in [2] and on previous work integrating trust, argumentation and belief revision [3], in this paper we presented a formal framework for modeling how different dimensions of the perceived trustworthiness of the source interact to determine the expected quality of the message, and how deviations from such expectation produce a specific feedback on source trustworthiness. In particular, we characterize competence and sincerity as the key ingredients of trustworthiness, and model both as being domain-dependent. Competence and trustworthiness determine an evaluation of trustworthiness for the source, which in turn produces an expectation on the quality of its arguments. Whenever an argument violates the recipient's expectation on quality, this produces a feedback on its source—positive if the argument was better than expected, negative if it was worse. Depending on several criteria (most notably, uncertainty reduction and social comparison), this

feedback is distributed among the two key dimensions of trustworthiness, to wit, competence and sincerity. Our model allows to detect the reasons behind the trustworthiness degree assigned to a source in a fine-gained way, and it is useful in such applications where the agents cannot simply avoid the interaction with the untrustworthy sources but they have to reason about trust.

Here, we applied this model to the case of agents exchanging and assessing arguments, but it could easily be extended to the exchange of any kind of factual information. The reason why we focused first on argumentation is because this provides a window on the agent's reasoning and the resulting process of belief change. This did not play a major role in the present paper, but it would be essential for most extensions of the model: for instance, to study how sources typically exchange information neither randomly nor out of mere kindness, but rather aiming at strategic changes in the recipient's beliefs and goals, to better serve the source's own agenda.

However, arguments in this paper were treated basically as black boxes, as it is most often the case in works based on abstract argumentation, in the vein of [6]. This is significant in two respects. First, we did not discuss the two-way relationship between source trustworthiness and trust in the message when what is being communicated is not the argument as a whole, but rather one of its constituents, e.g., a premise, its conclusion, or the inference rule licensing the argument, as in [13]. Finding out that the source is mistaken on the truth of some premise (hence the argument is unsound) rather than on the truth of the inference (hence the argument is invalid) is likely to have very different effects for the feedback on the source, which will have to be investigated in future work. Second, the internal structure of arguments in [3] is limited to deductively valid arguments, again as it is customary in abstract argumentation after Dung [6]. This is a huge idealization with respect to everyday argumentation: as informal logicians and argumentation theorists never tire to repeat [19], we rarely, if ever, exchange deductively valid arguments, while the vast majority of arguments are defeasible, which implies a different sort of consequence relation.

Finally, at present the framework does not capture the cumulative effect of converging sources on argument acceptability, or the effect of converging arguments on belief strength; instead, we use the maximum trustworthiness value among all the sources of an argument, and we do something similar for arguments converging on the same conclusion. This is acceptable in light of all the other issues addressed in this paper, and to comply with length constraints. However, when more than one source offer the same argument or piece of information, its acceptability is positively affected, even if the trustworthiness assigned to the additional sources is not especially high, as discussed in [1]. Conversely, the feedback from the message to the messenger might also depend on how many messengers were making that particular claim, and how much trusted each of them were to start with.

In spite of these limitations, our approach has the important merit of providing a unified framework to represent the effects of source trustworthiness on information assessment, and the converse impact of information quality on

source evaluation. This is the cornerstone of rational trust in communication: we assess neither the messenger nor its message in isolation, but instead capitalize on their mutual interdependence to obtain information on the trustworthiness of both. On this simple foundation, many future studies may (and ought to) build.

References

1. C. Castelfranchi. Representation and integration of multiple knowledge sources: issues and questions. In V. Cantoni, V. Di Gesù, A. Setti, and D. Tegolo, editors, *Human & Machine Perception: Information Fusion*. Plenum Press, 1997.
2. C. Castelfranchi and R. Falcone. *Trust Theory: A Socio-Cognitive and Computational Model*. Wiley, 2010.
3. C. da Costa Pereira, A. Tettamanzi, and S. Villata. Changing ones mind: Erase or rewind? In T. Walsh, editor, *IJCAI*, pages 164–171. IJCAI/AAAI, 2011.
4. J. Dix, S. Parsons, H. Prakken, and G. R. Simari. Research challenges for argumentation. *Computer Science - R&D*, 23(1):27–34, 2009.
5. D. Dubois and H. Prade. An introduction to bipolar representations of information and preference. *Int. J. Intell. Syst.*, 23:866–877, August 2008.
6. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.
7. K. Fullam and K. S. Barber. Using policies for information valuation to justify beliefs. In *AAMAS*, pages 404–411. IEEE Computer Society, 2004.
8. K. Fullam and K. S. Barber. Dynamically learning sources of trust information: experience vs. reputation. In *AAMAS*, page 164, 2007.
9. C.-J. Liau. Belief, information acquisition, and trust in multi-agent systems–a modal logic formulation. *Artif. Intell.*, 149(1):31–60, 2003.
10. E. Lorini and R. Demolombe. From binary trust to graded trust in information sources: A logical perspective. In *AAMAS-TRUST*, pages 205–225, 2008.
11. P.-A. Matt, M. Morge, and F. Toni. Combining statistics and arguments to compute trust. In *AAMAS*, pages 209–216, 2010.
12. S. Parsons, P. McBurney, and E. Sklar. Reasoning about trust using argumentation: A position paper. In *ArgMAS*, 2010.
13. S. Parsons, Y. Tang, E. Sklar, P. McBurney, and K. Cai. Argumentation-based reasoning in agents with varying degrees of trust. In *AAMAS*, pages 879–886, 2011.
14. H. Prade. A qualitative bipolar argumentative view of trust. In *SUM*, pages 268–276, 2007.
15. I. Rahwan and G. Simari, editors. *Argumentation in Artificial Intelligence*. Springer, 2009.
16. R. Stranders, M. de Weerdt, and C. Witteveen. Fuzzy argumentation for trust. In *CLIMA*, pages 214–230, 2007.
17. Y. Tang, K. Cai, E. Sklar, P. McBurney, and S. Parsons. A system of argumentation for reasoning about trust. In *EUMAS*, 2010.
18. S. Villata, G. Boella, D. M. Gabbay, and L. van der Torre. Arguing about the trustworthiness of the information sources. In *ECSQARU*, pages 74–85, 2011.
19. D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. CUP, 2008.
20. Y. Wang and M. P. Singh. Formal trust model for multiagent systems. In M. M. Veloso, editor, *IJCAI*, pages 1551–1556, 2007.
21. L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.