From trust in information sources to trust in communication systems: an analysis in modal logic

Emiliano Lorini and Robert Demolombe

Institut de Recherche en Informatique de Toulouse (IRIT), France
{lorini,demolombe}@irit.fr

Abstract. We present a logical analysis of trust that integrates in the definition of trust: the truster's goal and the truster's belief that the trustee has the right properties (powers, abilities, dispositions) to ensure that the goal will be achieved. The second part of the paper is focused on the specific domain of trust in information sources and communication systems. We provide an analysis of the properties of information sources (validity, completeness, sincerity, competence, vigilance and cooperativity) and communication systems (availability and privacy) and, we discuss their relationships with trust.

1 Introduction

Future computer applications such as the semantic Web [1], e-business and e-commerce [13], Web services [24] will be open distributed systems in which the many constituent components are agents spread throughout a network in a decentralized manner. These agents will interact between them in flexible ways in order to achieve their design objectives and to accomplish the tasks which are delegated to them by the human users. Some of them will directly interact and communicate with the human users. During the system's lifetime, these agents will need to manage and deal with trust. They will need to automatically make trust judgments in order to assess the trustworthiness of other (software and human) agents while, for example, exchanging money for a service, giving access to a certain information, choosing between conflicting sources of information. They will also need to understand how trust can be induced in a human user in order to support his interaction with the system and to motivate him to use the application. Consequently, these agents will need to understand the components and the determinants of the user's trust in the system.

Thus, to realize all their potential, future computer applications will require the development of sophisticated formal and computational models of trust. These models must provide clear definitions of the relevant concepts related to trust and safe reasoning rules which can be exploited by the agents for assessing the trustworthiness of a given target. Moreover, these models of trust must be cognitively plausible, so that they can be directly exploited by the agents during their interactions with the human user in order to induce him to trust the system and the underlying Information and Communication Technology (ICT) infrastructure. With cognitively plausible models of trust, we mean models in which the main cognitive constituents of trust as a mental attitude are identified (e.g. beliefs, goals).

This paper follows our previous works [21, 16] with the objective of developing a general formal model of trust which meets the previous desiderata. It is worth noting that it is not our aim to propose a model of trust based on statistics about past interactions with a given target and reputational information. In particular, the present paper focuses on an issue that we have neglected up to now: the issue of trust in information sources and communication systems. We think that this issue is very relevant for future computer applications such as the semantic Web, e-business and Web services. For example, in a typical scenario of e-business, trust in information sources has a strong influence on an agent's decision to buy, or to sale, a specific kind of stocks. Indeed, to take such a decision an agent has several types of information sources to consult in order to predict the future evolution of the stock value. These information sources may be banks, companies, consultants, *etc*. and the agent may believe that some of these information sources have a good competence but are not necessarily sincere, others are reluctant to inform about bad news, others are competent but are not necessarily informed at the right moment, *etc*. In a typical scenario of Web services, an agent might want to make a credit card transaction by means of a certain online payment system. In this case, the agent's trust in the communication system has a strong influence on the agent's decision to exploit it for the credit card transaction. In particular, the agent's trust in the online payment system is supported by the agent's belief that the online payment system will ensure the privacy of the credit card number from potential intruders.

The paper is organized as follows. We start with a presentation of a modal logic which enables reasoning about actions, beliefs and goals of agents (Section 2). This logic will be used during the paper for formalizing the relevant concepts of our model of trust. Then, a general definition of trust is presented (Section 3). Section 4 is focused on the formal characterization of the main properties of an information source: validity, completeness, sincerity, competence, vigilance and cooperativity. In Section 5 we show that these properties are epistemic supports for trust in information sources. In section 6 we provide an analysis of communication systems. We define two fundamental properties of communication systems: availability and privacy. Then, in Section 7, we show that these properties are epistemic supports for an agent's trust in a communication system. We conclude with a discussion of some related works and we show some directions for future works.

2 A modal logic of beliefs, goals and actions

We present in this section the multimodal logic $\mathcal{L}$ that we use in the paper to formalize the relevant concepts of our model of trust. $\mathcal{L}$ combines the expressiveness of a dynamic logic [15] with the expressiveness of a logic of agents' mental attitudes [7, 25].

2.1 Syntax and semantics

The syntactic primitives of the logic $\mathcal{L}$ are the following: a nonempty finite set of agents $AGT = \{i, j, \ldots\}$; a nonempty finite set of atomic actions $AT = \{a, b, \ldots\}$; a finite set of atomic formulas $\Pi = \{p, q, \ldots\}$. LIT is the set of literals which includes all atomic formulas and their negations, that is, $LIT = \{p, \neg p | p \in \Pi\}$. We note $P, Q, \ldots$ the elements in LIT. We also introduce specific actions of the form

$inf_j(P)$ denoting the action of informing agent j that P is true. We call them informative actions. The set INFO of informative actions is defined as follows: $\mathit{INFO} = \{inf_j(P) | j \in AGT,\ P \in LIT\}$. Since the set Π is finite, the set INFO is finite as well. The set ACT of complex actions is given by the union of the set of atomic actions and the set of informative actions, that is: $ACT = AT \cup \mathit{INFO}$. We note $\alpha, \beta, \ldots$ the elements in ACT. The language of $\mathcal{L}$ is the set of formulas defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid After_{i:\alpha}\varphi \mid Does_{i:\alpha}\varphi \mid Bel_i\varphi \mid Goal_i\varphi$$

where p ranges over Π, α ranges over ACT and i ranges over AGT. The operators of our logic have the following intuitive meaning. $Bel_i\varphi$: the agent i believes that φ; $After_{i:\alpha}\varphi$: after agent i does α, it is the case that φ ($After_{i:\alpha}\bot$ is read: agent i cannot do action α); $Does_{i:\alpha}\varphi$: agent i is going to do α and φ will be true afterward ($Does_{i:\alpha}\top$ is read: agent i is going to do α); $Goal_i\varphi$: the agent i wants that φ holds. The following abbreviations are given: $Can_i(\alpha) \stackrel{\text{def}}{=} \neg After_{i:\alpha}\bot$;
$Int_i(\alpha) \stackrel{\text{def}}{=} Goal_i Does_{i:\alpha}\top$;
$Inf_{i,j}(P) \stackrel{\text{def}}{=} Does_{i:inf_j(P)}\top$.
$Can_i(\alpha)$ stands for: agent i can do action α (i.e. i has the capacity to do α). $Int_i(\alpha)$ stands for: agent i intends to do α. Finally $Inf_{i,j}(P)$ stands for: i informs j that P is true. Models of the logic $\mathcal{L}$ are tuples $M = \langle W, R, D, B, G, V\rangle$ defined as follows.

- W is a non empty set of possible worlds or states.
- $R : AGT \times ACT \longrightarrow W \times W$ maps every agent i and action α to a relation $R_{i:\alpha}$ between possible worlds in W. Given a world $w \in W$, if $(w, w') \in R_{i:\alpha}$ then w' is a world which can be reached from w through the occurrence of agent i's action α.
- $D : AGT \times ACT \longrightarrow W \times W$ maps every agent i and action α to a relation $D_{i:\alpha}$ between possible worlds in W. Given a world $w \in W$, if $(w, w') \in D_{i:\alpha}$ then w' is the **unique actual RW1 RW3** *next* world of w which will be reached from w through the occurrence of agent i's action α.
- $B : AGT \longrightarrow W \times W$ maps every agent i to a serial, transitive and euclidean relation B_i between possible worlds in W. Given a world $w \in W$, if $(w, w') \in B_i$ then w' is a world which is compatible with agent i's beliefs at w.
- $G : AGT \longrightarrow W \times W$ maps every agent i to a serial relation G_i between possible worlds in W. Given a world $w \in W$, if $(w, w') \in G_i$ then w' is a world which is compatible with agent i's goals at w.
- $V : W \longrightarrow 2^{\Pi}$ is a truth assignment which associates each world w with the set $V(w)$ of atomic propositions true in w.

Given a model M, a world w and a formula φ, we write $M, w \models \varphi$ to mean that φ is true at world w in M, under the basic semantics. The rules defining the truth conditions of formulas are just standard for atomic formulas, negation and disjunction. The following are the remaining truth conditions for $After_{i:\alpha}\varphi$, $Does_{i:\alpha}\varphi$, $Bel_i\varphi$ and $Goal_i\varphi$.

- $M, w \models After_{i:\alpha}\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in R_{i:\alpha}$
- $M, w \models Does_{i:\alpha}\varphi$ iff $\exists w'$ such that $(w, w') \in D_{i:\alpha}$ and $M, w' \models \varphi$
- $M, w \models Bel_i\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in B_i$
- $M, w \models Goal_i\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in G_i$

The following section is devoted to illustrate the additional semantic constraints over $\mathcal{L}$ models and the corresponding axiomatization of the logic $\mathcal{L}$.

2.2 Axiomatization

Operators for actions of type $After_{i:\alpha}$ and $Does_{i:\alpha}$ are supposed to be normal modal operators satisfying the axioms and rules of inference of system K. Operators for belief of type Bel_i are supposed to be $KD45$ normal modal operators, whilst operators for goal of type $Goal_i$ are supposed to be KD normal modal operators. Thus, we make assumptions about positive and negative introspection for beliefs and we suppose that an agent have no inconsistent beliefs or conflicting goals.

We add the following constraint over every relation $D_{i:\alpha}$ and every relation $D_{j:\beta}$ of all $\mathcal{L}$ models. For every $i, j \in AGT$, $\alpha, \beta \in ACT$ and $w \in W$:

S1 if $(w, w') \in D_{i:\alpha}$ and $(w, w'') \in D_{j:\beta}$ then $w' = w''$

Constraint $S1$ says that if w' is the *next* world of w which is reachable from w through the occurrence of agent i's action α and w'' is also the *next* world of w which is reachable from w through the occurrence of agent j's action β, then w' and w'' denote the same world. Indeed, we suppose that every world can only have one *next* world. The semantic constraint $S1$ corresponds to the following axiom.

Alt$_{Act}$ $Does_{i:\alpha}\varphi \rightarrow \neg Does_{j:\beta}\neg\varphi$

Axiom **Alt**$_{Act}$ says that: if i is going to do α and φ will be true afterward, then it cannot be the case that j is going to do β and $\neg\varphi$ will be true afterward.

We also suppose that the world is never static in our framework, that is, we suppose that for every world w there exists some agent i and action α such that i is going to perform α at w. Formally, for every $w \in W$ we have that:

S2 $\exists i \in AGT, \exists \alpha \in ACT, \exists w' \in W$ such that $(w, w') \in D_{i:\alpha}$

The semantic constraint $S2$ corresponds to the following axiom of our logic.

Active $\bigvee_{i \in AGT, \alpha \in ACT} Does_{i:\alpha}\top$

Axiom **Active** ensures that for every world w there is a *next* world of w which is reachable from w by the occurrence of some action of some agent. This is the reason why the operator X for *next* of LTL (linear temporal logic) can be defined as follows:[1]

$$X\varphi \stackrel{\text{def}}{=} \bigvee_{i \in AGT, \alpha \in ACT} Does_{i:\alpha}\varphi$$

The following relationship is supposed between every relation $D_{i:\alpha}$ and the corresponding relation $R_{i:\alpha}$ of all $\mathcal{L}$ models. For every $i \in AGT$, $\alpha \in ACT$ and $w \in W$:

S3 if $(w, w') \in D_{i:\alpha}$ then $(w, w') \in R_{i:\alpha}$

The constraint $S3$ says that if w' is the *next* world of w which is reachable from w through the occurrence of agent i's action α, then w' is a world which is *possibly* reachable from w through the occurrence of agent i's action α. The semantic constraint $S3$ corresponds to the following axiom **Inc**$_{Act,PAct}$.

[1] Note that X satisfies the standard property $X\varphi \leftrightarrow \neg X \neg\varphi$ (i.e. φ will be true in the next state iff $\neg\varphi$ will not be true in the next state).

Inc$_{Act,PAct}$ $Does_{i:\alpha}\varphi \rightarrow \neg After_{i:\alpha}\neg\varphi$

According to **Inc**$_{Act,PAct}$, if i is going to do α and φ will be true afterward, then it is not the case that $\neg\varphi$ will be true after i does α. The following axioms relates intentions with actions.

IntAct1 $(Int_i(\alpha) \wedge Can_i(\alpha)) \rightarrow Does_{i:\alpha}\top$
IntAct2 $Does_{i:\alpha}\top \rightarrow Int_i(\alpha)$

According to **IntAct1**, if i has the intention to do action α and has the capacity to do α, then i is going to do α. According to **IntAct2**, an agent is going to do action α only if he has the intention to do α. In this sense we suppose that an agent's *doing* is by definition intentional. Similar axioms have been studied in [22] in which a logical model of the relationships between intention and action performance is proposed. **IntAct1** and **IntAct2** correspond to the following semantic constraints over $\mathcal{L}$ models. For every $i \in AGT$, $\alpha \in ACT$ and $w \in W$:

S4 if $\forall(w, w') \in G_i, \exists w''$ such that $(w', w'') \in D_{i:\alpha}$ and $\exists v$ such that $(w, v) \in R_{i:\alpha}$ then $\exists v'$ such that $(w, v') \in D_{i:\alpha}$
S5 if $\exists v'$ such that $(w, v') \in D_{i:\alpha}$ then $\forall(w, w') \in G_i, \exists w''$ such that $(w', w'') \in D_{i:\alpha}$

As far as informative actions are concerned, we assume that they are always executable, *i.e.*an agent i can always inform another agent j about a fact P. Formally:

CanInf $Can_i(inf_j(P))$

Axiom **CanInf** corresponds to the following semantic constraint over $\mathcal{L}$ models. For every $i \in AGT$, $inf_j(P) \in INFO$ and $w \in W$:

S6 $\exists w'$ such that $(w, w') \in R_{i:inf_j(P)}$

We also suppose that goals and beliefs must be compatible, that is, if an agent has the goal that φ then, he cannot believe that $\neg\varphi$. Indeed, the notion of goal we characterize here is a notion of an agent's *chosen goal*, i.e. a goal that an agent decides to pursue. As some authors have stressed (e.g. [3]), a rational agent cannot decide to pursue a certain state of affairs φ, if he believes that $\neg\varphi$. Thus, for any $i \in AGT$ and $w \in W$ the following semantic constraint over $\mathcal{L}$ models is supposed:

S7 $\exists w'$ such that $(w, w') \in B_i$ and $(w, w') \in G_i$

The constraint $S7$ corresponds to the following axiom **WR** (*weak realism*) of our logic.

WR $Goal_i\varphi \rightarrow \neg Bel_i\neg\varphi$

In this work we assume positive and negative introspection over (chosen) goals, that is:

PIntrGoal $Goal_i\varphi \rightarrow Bel_iGoal_i\varphi$
NIntrGoal $\neg Goal_i\varphi \rightarrow Bel_i\neg Goal_i\varphi$

Axioms **PIntrGoal** and **NIntrGoal** correspond to the following semantic constraints over $\mathcal{L}$ models. For any $i \in AGT$ and $w \in W$:

S8 if $(w, w') \in B_i$ then $\forall v$, if $(w, v) \in G_i$ then $(w', v) \in G_i$
S9 if $(w, w') \in B_i$ then $\forall v$, if $(w', v) \in G_i$ then $(w, v) \in G_i$

We suppose that agents satisfy the property of *no forgetting* (**NF**)[2], that is, if an agent i believes that after agent j does α, it is the case that φ, and agent i does not believe that j cannot do action α, then after agent j does α, i believes that φ.

NF $(Bel_i After_{j:\alpha}\varphi \wedge \neg Bel_i \neg Can_j(\alpha)) \rightarrow After_{j:\alpha} Bel_i \varphi$

Axiom **NF** corresponds to the following semantic constraint over $\mathcal{L}$ models. For any $i, j \in AGT$, $\alpha \in ACT$, and $w \in W$:

S10 if $(w, w') \in R_{j:\alpha} \circ B_i$ and $\exists v$ such that $(w, v) \in B_i \circ R_{j:\alpha}$ then $(w, w') \in B_i \circ R_{j:\alpha}$

where $\circ$ is the standard composition operator between two binary relations. In accepting the axiom **NF**, we suppose that events are always uninformative, that is, i should not forget anything about the particular effects of j's action α that starts at a world w. What an agent i believes at a world w', only depends on what i believed at the previous world w and on the action which has occurred and which was responsible for the transition from w to w'. Besides, the axiom **NF** relies on an additional assumption of complete and correct information. It is supposed that j's action α occurs if and only if every agent is informed of this fact. Hence all action occurrences are supposed to be public.

We also have specific properties for informative actions. We suppose that if an agent i is informed (resp. not informed) by another j that some fact P is true then i is aware of being informed (resp. not being informed) by j.

PIntrInf $Inf_{j,i}(P) \rightarrow Bel_i Inf_{j,i}(P)$
NIntrInf $\neg Inf_{j,i}(P) \rightarrow Bel_i \neg Inf_{j,i}(P)$

Axioms **PIntrInf** and **NIntrInf** correspond to the following semantic constraints over $\mathcal{L}$ models. For any $i, j \in AGT$, $inf_i(P) \in INFO$, and $w \in W$:

S11 if $\exists w'$ sutch that $(w, w') \in D_{j:inf_i(P)}$ then $\forall (w, v) \in B_i, \exists w''$ such that $(v, w'') \in D_{j:inf_i(P)}$
S12 if $\exists w', w''$ such that $(w, w') \in B_i$ and $(w', w'') \in D_{j:inf_i(P)}$ then $\exists v$ sutch that $(w, v) \in D_{j:inf_i(P)}$

We call $\mathcal{L}$ the logic axiomatized by the axioms and rules of inference presented above. We write $\vdash \varphi$ if formula φ is a theorem of $\mathcal{L}$ (i.e. φ is the derivable from the axioms and rules of inference of the logic $\mathcal{L}$). We write $\models \varphi$ if φ is *valid* in all $\mathcal{L}$ models, i.e. $M, w \models \varphi$ for every $\mathcal{L}$ model M and world w in M. Finally, we say that φ is *satisfiable* if there exists a $\mathcal{L}$ model M and world w in M such that $M, w \models \varphi$. We can prove that the logic $\mathcal{L}$ is *sound* and *complete* with respect to the class of $\mathcal{L}$ models. Namely:

Theorem 1 $\vdash \varphi$ *if and only if* $\models \varphi$.

Proof 1 *It is a routine task to check that all the axioms of the logic $\mathcal{L}$ correspond to their semantic counterparts. It is routine, too, to check that all of our axioms are in the Sahlqvist class, for which a general completeness result exists [2].*

[2] See also [12, 27] for a discussion of this property.

3 A general definition of trust

In this work trust is conceived as a complex configuration of mental states in which there is both a motivational component and an epistemic component. More precisely, we assume that an agent i's trust in agent j necessarily involves a goal of the truster: if agent i trusts agent j then, necessarily, i trusts j with respect to some of his goals. The core of trust is a belief of the truster about some properties of the trustee, that is, if agent i trusts agent j then necessarily i trusts j because i has some goal and believes that j has the right properties to ensure that such a goal will be achieved. The concept of trust formalized in this work is similar to the concept of trust defined by Castelfranchi & Falcone [6, 5]. We agree with them that trust should not be seen as an unitary and simplistic notion as other models implicitly suppose. For instance, there are computational models of trust in which trust is conceived as an expectation of the truster about a successful performance of the trustee sustained by the repeated direct interactions with the trustee (under the assumption that iterated experiences of success strengthen the truster's confidence) [19]. More sophisticated models of social trust have been developed in which reputational information is added to information obtained via direct interaction (e.g. [17, 26]). All these models are in our view over-simplified since they do not consider the beliefs supporting the truster's evaluation of the trustee.

On this point we agree with Castelfranchi & Falcone on the fact that trust is based on the truster's *evaluation* of specific properties of the trustee (e.g. abilities, competencies, dispositions, etc.) and of the environment in which the trustee is going to act, which are relevant for the achievement of a goal of the truster. From this perspective, trust is nothing more than the truster's belief about some relevant properties of the trustee with respect to a given goal. [3] The following is the concept of trust as an *evaluation* that interests us in this paper.

Definition 1 ***TRUST IN THE TRUSTEE'S ACTION***. *i trusts j to do α with regard to his goal that φ if and only if i wants φ to be true and i believes that:*[4]

1. *j, by doing α, will ensure that φ AND*
2. *j has the capacity to do α AND*
3. *j intends to do α*

The formal translation of Definition 1 is: [5] [6]

$$Trust(i, j, \alpha, \varphi) \stackrel{\text{def}}{=} Goal_i X\varphi \wedge Bel_i(After_{j:\alpha}\varphi \wedge Can_j(\alpha) \wedge Int_j(\alpha))$$

In our logic the conditions $Can_j(\alpha)$ and $Int_j(\alpha)$ together are equivalent to $Does_{j:\alpha}\top$ (by axioms $\mathbf{Inc}_{Act,PAct}$, **IntAct1** and **IntAct2**), so the definition of trust in the trustee's action can be simplified as follows:

[3] In this paper we do not consider a related notion of *decision to trust*, that is, the truster's decision to bet and wager on the trustee and to rely on her for the accomplishment of a given task. For a distinction between trust as an *evaluation* and trust as a *decision*, see [6, 23].

[4] In the present paper we only focus on *full trust* involving a *certain belief* of the truster. In order to extend the present analysis to forms of *partial trust*, a notion of *graded belief* (i.e. uncertain belief) or *graded trust*, as in [11], is needed.

[5] **Notice that positive and negative introspection of trust follows from similar properties for beliefs and goals. RW1**

[6] **The meaning of the goal φ may be that sometimes in the future ψ holds. RW3.**

$$Trust(i,j,\alpha,\varphi) \stackrel{\text{def}}{=} Goal_i X\varphi \wedge Bel_i(After_{j:\alpha}\varphi \wedge Does_{j:\alpha}\top)$$

$Trust(i,j,\alpha,\varphi)$ is meant to stand for: i trusts j to do α with regard to his goal that φ.

Example 1. The two agents i and j are making a transaction in Internet. After having paid j, i trusts j to send him a certain product with regard to his goal of having the product in the next state:

$$Trust(i,j,send,HasProduct(i)).$$

This means that i wants to have the product in the next state:

$$Goal_i X HasProduct(i).$$

Moreover, according to i's beliefs, j, by sending him the product, will ensure that he will have the product in the next state, and j is going to send the product:

$$Bel_i(After_{j:send}HasProduct(i) \wedge Does_{j:send}\top).$$

The following theorem highlights the fact that if i trusts j to do α with regard to his goal that φ then i has a positive expectation that φ will be true in the next state.

Theorem 2 *Let* $i,j \in AGT$ *and* $\alpha \in ACT$. *Then:*
$\vdash Trust(i,j,\alpha,\varphi) \rightarrow Bel_i X\varphi$

In our view *trust in the trustee's action* must be distinguished from *trust in the trustee's inaction*. The former is focused on the domain of gains (goal achievements) whereas the latter is focused on the domain of losses (goal frustrations). That is, in the former case the truster believes that the trustee is in condition to *further* the achievement of his goals, and she will do that; in the latter case the truster believes that the trustee is in condition to *endanger* the achievement of his goals, but she will not do that. Trust in the trustee's inaction is based on the fact that, by doing some action α, agent j can prevent i to reach his goal. In that case i expects that j will not intend to do α.

Definition 2 ***TRUST IN THE TRUSTEE'S INACTION***. *i trusts j not to do α with regard to his goal φ if and only if i wants φ to be true and i believes that:*

1. *j, by doing α, will ensure that $\neg\varphi$ AND*
2. *j has the capacity to do α AND*
3. *j does not intend to do α*

The formal definition of trust in the trustee's inaction is given by the following abbreviation.

$$Trust(i,j,\neg\alpha,\varphi) \stackrel{\text{def}}{=} Goal_i X\varphi \wedge Bel_i(After_{j:\alpha}\neg\varphi \wedge Can_j(\alpha) \wedge \neg Int_j(\alpha))$$

$Trust(i,j,\neg\alpha,\varphi)$ stands for: i trusts j not to do α with regard to his goal that φ.

Example 2. Agent j is the webmaster of a public access website with financial information. Agent i is a regular reader of this website and he trusts j not to restrict the access to the website with regard to his goal of having free access to the website:

$$Trust(i,j,\neg restrict, freeAccess(i)).$$

This means that, i has the goal of having free access to the website in the next state:

$$Goal_i X freeAccess(i).$$

Moreover, according to i's beliefs, j has the capacity to restrict the access to the website and, by restricting the access to the website, j will ensure that i will not have free access to the website in the next state, but j does not intend to restrict the access:

$$Bel_i(After_{j:restrict}\neg freeAccess(i) \wedge Can_j(restrict) \wedge \neg Int_j(restrict)).$$

In this situation, i's trust in j is based on i's belief that j is in condition to restrict the access to the website, but she does not have the intention to do this.

Note that, differently from agent i's trust in agent j's action, agent i's trust in agent j's inaction with respect to the goal that φ does not entail i's positive expectation that φ will be true. Indeed, $Trust(i, j, \neg\alpha, \varphi) \wedge \neg Bel_i X\varphi$ is satisfiable in our logic. The intuitive reason is that $\neg\varphi$ may be the effect of another action than $j : \alpha$.

In the following sections 4 and 5 we will study the properties of information sources and show how these properties can be evaluated by the truster in order to assess the trustworthiness of an information source.

4 Basic properties of an information source

We suppose that the properties of an information source can be defined in terms of the relationships between three facts: an information source j informs an agent i that a certain fact P is true; an information source j believes that P is true; the fact P is true. These properties are all expressed in a conditional form. Since we have three facts that can be related in a similar conditional form, a systematic analysis of these relationships leads to six different properties of information sources.

Definition 3 ***INFORMATION SOURCE VALIDITY***. *Agent j is a valid information source about P with regard to i if and only if, after j does the action of informing i about P, it is the case that P.*

Formally: $Valid(j, i, P) \stackrel{\text{def}}{=} After_{j:inf_i(P)}P$

Note $After_{j:inf_i(P)}P$ can also be read in an explicit conditional form: if j does the action $inf_i(P)$, then P is true after the action has been done.

Definition 4 ***INFORMATION SOURCE COMPLETENESS***. *Agent j is a complete information source about P with regard to i if and only if, if P is true then j does the action of informing i about P.*

Formally: $Compl(j, i, P) \stackrel{\text{def}}{=} P \rightarrow Inf_{j,i}(P)$

Definition 5 ***INFORMATION SOURCE SINCERITY***. *Agent j is a sincere information source about P with regard to i if and only if, if j does the action of informing i about P then j believes that P.*

Formally: $Sinc(j, i, P) \stackrel{\text{def}}{=} Inf_{j,i}(P) \rightarrow Bel_j P$

Definition 6 ***INFORMATION SOURCE COMPETENCE***. *Agent j is a competent information source about P if and only if, if j believes P then P is true.*

Formally: $Compet(j, P) \stackrel{\text{def}}{=} Bel_j P \rightarrow P$

Definition 7 ***INFORMATION SOURCE VIGILANCE***. *Agent j is a vigilant information source about P if and only if, if P is true then j believes P.*

Formally: $Vigil(j, P) \stackrel{\text{def}}{=} P \rightarrow Bel_j P$

Definition 8 ***INFORMATION SOURCE COOPERATIVITY***. *Agent j is a cooperative information source about P with regard to i if and only if, if j believes that P then j informs i about P.* [7]

Formally: $Coop(j, i, P) \stackrel{\text{def}}{=} Bel_j P \rightarrow Inf_{j,i}(P)$

It is worth noting that the previous properties of information sources are not independent. For instance, as the following theorem 3 shows, validity can be derived from sincerity and competence and, completeness from vigilance and cooperativity.

Theorem 3 *Let $i, j \in AGT$ and $inf_i(P) \in$ INFO, then:*

1. $\vdash After_{j:inf_i(P)}(Compet(j, P) \wedge Sinc(j, i, P)) \rightarrow Valid(j, i, P)$
2. $\vdash (Vigil(j, P) \wedge Coop(j, i, P)) \rightarrow Compl(j, i, P)$

Note that in theorem 3.1, the derivation of $Valid(j, i, P)$ requires j's competence and sincerity at the instant where the action $inf_i(P)$ has been done. This is the reason why we have $After_{j:inf_i(P)}(Compet(j, P) \wedge Sinc(j, i, P))$ in the antecedent.

Example 3. Consider an example in the field of stocks and bonds market. The agent BUG is the Bank of Union of Groenland. Sue Naive (SN) and Very Wise (VW) are two BUG's customers. BUG plays the role of an information source for the customers, for instance for the facts p: "it is recommended to buy MicroHard stocks", and q: "Microhard stocks are dropping". SN believes that BUG is sincere with regard to her about p and BUG is competent about p, because SN believes that BUG wants to help its customers and BUG has a long experience in the domain. SN also believes that BUG is cooperative with regard to her about q because q is a relevant information for customers in order to make decisions. VW too believes that BUG is competent about p. But VW does not believe that BUG is sincere with regard to him about p. Indeed, VW believes that BUG wants that VW buys Microhard stocks, even if this is not profitable for VW. This example is formally represented by the following formula:

$$Bel_{SN}Sinc(BUG, SN, p) \wedge Bel_{SN}Compet(BUG, p) \wedge Bel_{SN}Coop(BUG, SN, q) \wedge$$
$$Bel_{VW}Compet(BUG, p) \wedge \neg Bel_{VW}Sinc(BUG, VW, p).$$

5 Trust in information sources

We conceive trust in information sources as a specific instance of the general notion of trust in the trustee's action defined in section 3. In our view, the relevant aspect of trust in information sources is the content of the truster's goal. In particular, we suppose that an agent i trusts the information source j to inform him whether the fact P is true only if i has the *epistemic goal* of knowing whether P is true and believes that, due to the information transmitted by j, he will achieve this goal. In this sense, trust in information sources is characterized by an epistemic goal of the truster and an informative action of the trustee. The concept of epistemic goal can be defined from the following standard definitions of *knowing that* (i.e. as having the correct belief that something is the case) and *knowing whether*:

$$K_i\varphi \stackrel{\text{def}}{=} Bel_i\varphi \wedge \varphi \quad KW_i\varphi \stackrel{\text{def}}{=} K_i\varphi \vee K_i\neg\varphi$$

[7] **This definition of cooperativity does not exclude that i does not want to be informed about p, like in spamming. RW3.**

where $K_i\varphi$ stands for "agent i knows that φ is true" and, $KW_i\varphi$ stands for "i knows whether φ is true". An *epistemic goal* of an agent i is i's goal of knowing the truth value of a certain formula. Formally, $Goal_i KW_i\varphi$ denotes i's epistemic goal of knowing whether φ is true now; $Goal_i X\ KW_i\varphi$ denotes i's epistemic goal of knowing whether φ is true in the next state.

Our aim in this section of the paper is to investigate the relationships between trust in information sources and the properties of information sources defined in section 4. The following theorem 4 highlights the relationship between trust in information sources and the properties of validity and completeness of information sources. It says that: if i believes that j is a valid information source about p and $\neg p$ with regard to i and that j is a complete information source about p and $\neg p$ with regard to i and, i has the epistemic goal of knowing whether p is true then, either i trusts the information source j to inform him that p is true or i trusts the information source j to inform him that $\neg p$ is true (with regard to his epistemic goal of knowing whether p is true).

Theorem 4 *Let $i, j \in AGT$ and $inf_i(p), inf_i(\neg p) \in$ INFO, then:*
$\vdash (Bel_i(Valid(j,i,p) \wedge Valid(j,i,\neg p)) \wedge Bel_i(Compl(j,i,p) \wedge Compl(j,i,\neg p)) \wedge Goal_i X\ KW_i p) \rightarrow (Trust(i,j,inf_i(p),KW_i p) \vee Trust(i,j,inf_i(\neg p),KW_i p))$

The reason why in the consequent we have a disjunction (instead of a conjunction) is that p is either true or false. RW1. Then, j may inform i either about the truth of p or about the truth of $\neg p$. From theorems 3.1 and 3.2, similar theorems can be proved by substituting $Valid(j,i,p)$ with $After_{j:inf_i(p)}(Compet(j,p) \wedge Sinc(j,i,p))$, $Valid(j,i,\neg p)$ with $After_{j:inf_i(\neg p)}$ $(Compet(j,\neg p) \wedge Sinc(j,i,\neg p))$, $Compl(j,i,p)$ with $Vigil(j,p) \wedge Coop(j,i,p)$ and, $Compl(j,i,\neg p)$ with $Vigil(j,\neg p) \wedge Coop(j,i,\neg p)$ in the antecedent of theorem 4.

The following theorem 5 is a specific instantiation of theorem 2. It says that: if i trusts the information source j to inform him that p or i trusts the information source j to inform him that $\neg p$ with regard to his goal of knowing whether p is true, then i believes that in the next state he will achieve his goal of knowing whether p is true.

Theorem 5 *Let $i, j \in AGT$ and $inf_i(p), inf_i(\neg p) \in$ INFO, then:*
$\vdash (Trust(i,j,inf_i(p),KW_i p) \vee Trust(i,j,inf_i(\neg p),KW_i p)) \rightarrow Bel_i X\ KW_i p$

Example 4. Let us consider again the example of stocks and bonds market. SN has the epistemic goal of knowing whether q ("Microhard stocks are dropping") is true:

$$Goal_{SN} X\ KW_{SN} q.$$

SN believes that BUG is a valid information source with regard to her both about q and about $\neg q$ and that BUG is a complete information source with regard to her both about q and about $\neg q$:

$$Bel_{SN}(Valid(BUG,SN,q) \wedge Valid(BUG,SN,\neg q)) \wedge$$
$$Bel_{SN}(Compl(BUG,SN,q) \wedge Compl(BUG,SN,\neg q)).$$

Then, by theorem 4, we can infer that either SN trusts the information source BUG to inform her that q is true or SN trusts the information source BUG to inform her that $\neg q$ is true (with regard to her epistemic goal of knowing whether q is true):

$$Trust(SN,BUG,inf_{SN}(q),KW_{SN} q) \vee Trust(SN,BUG,inf_{SN}(\neg q),KW_{SN} q).$$

Finally, by theorem 5, we can infer that SN believes that in the next state she will achieve her goal of knowing whether q is true:

$$Bel_{SN}X\ KW_{SN}q.$$

In the following two sections 6 and 7 we will shift the focus of analysis from information sources to communication systems. We will study some important properties of communication systems and show how these properties can be evaluated by the truster in order to assess the trustworthiness of a communication system.

6 Basic properties of a communication system

We suppose that the fundamental properties of a communication system j can be defined in terms of two facts: the communication system j satisfies an agent i's goal that a certain information will be transmitted to another agent z or, the communication system j satisfies an agent i's goal that a certain information will not be transmitted to another agent z. In the former case we say that the communication system j is available to i to transmit the information. In the latter case we say that the communication system j ensures to i the privacy of the information. For simplification, we ignore in this work other properties of communication systems like authentication and integrity.

Definition 9 ***COMMUNICATION SYSTEM AVAILABILITY***. *The communication system j is available to agent i to transmit the information P to agent z if and only if, if j believes that i wants that j informs z about P then j informs z about P.*

Formally: $Avail(j,i,z,P) \stackrel{\text{def}}{=} Bel_j Goal_i Inf_{j,z}(P) \rightarrow Inf_{j,z}(P)$

Definition 10 ***COMMUNICATION SYSTEM PRIVACY***. *The communication system j ensures to agent i the privacy of information P from agent z if and only if, if j believes that i does not want that j informs z about P then, j does not inform z about P.*

Formally: $Priv(j,i,z,P) \stackrel{\text{def}}{=} Bel_j Goal_i \neg Inf_{j,z}(P) \rightarrow \neg Inf_{j,z}(P)$

Example 5. Let us consider an example in the field of Web services. An agent called Bill decides to use a Hotel Booking Service (HBS) in Internet in order to book a double room at the Hotel Colosseum (HC) in Rome. Bill's decision is affected by two beliefs of Bill, the belief that HBS ensures the privacy from a potential intruder of the information r: "Bill's credit card number is 01234567891234", the belief that HBS is available to inform HC about s: "Bill has made an online reservation". According to our definitions, this example is formally represented by:

$$Bel_{Bill}Priv(HBS, Bill, intruder, r) \wedge Bel_{Bill}Avail(HBS, Bill, HC, s).$$

7 Trust in communication systems

We conceive trust in communication systems as a specific instance of the notion of trust defined in section 3. On the one hand, we suppose that an agent i trusts the communication system j to inform agent z about P with regard to his goal that z believes P (*trust in a communication system's action*) if and only if, i has the goal that z believes P and i believes that, due to the information transmitted by j to z, z will believe P. On the other hand, we suppose that an agent i trusts the communication system j not to inform agent z about P with regard to his goal that z does not believe P (*trust in a* *communication system's inaction*) if and only if, i has the goal that z does not believe P, i believes that, by informing z about P, j will ensure that z believes P but, i believes that j does not intend to inform z about P. In this sense, i's trust in the communication system j's action (resp. inaction) is characterized by i's goal that a certain information will be transmitted (resp. will not be transmitted) to another agent z so that z will have access (resp. will not have access) to this information.

Our aim in this section of the paper is to investigate the relationships between trust in communication systems and the two properties of communication systems defined in section 6. The following theorem 6 highlights the relationship between trust in a communication system's action and the availability of the communication system. It says that: if i has the goal that in the next state z will believe P, i believes that j is available to inform z about P, i believes that j believes that i wants j to inform z about P and, i believes that z believes that j is a valid information source about P then, i trusts j to inform z about P with regard to his goal that z will believe P.

Theorem 6 *Let $i, j, z \in AGT$ and $inf_z(P) \in$ INFO, then:*
$\vdash (Goal_i X Bel_z P \wedge Bel_i Avail(j, i, z, P) \wedge Bel_i Bel_j Goal_i Inf_{j,z}(P) \wedge$
$Bel_i Bel_z Valid(j, z, P)) \rightarrow Trust(i, j, inf_z(P), Bel_z P)$

Example 6. Let us consider again the example of the Hotel Booking Service (HBS). Bill has the goal that the receptionist at the Hotel Colosseum (HC) believes that s ("Bill has made an online reservation"):

$$Goal_{Bill} X Bel_{HC} s.$$

Bill believes that HBS is available to inform the HC's receptionist about s and that HBS believes that Bill wants HBS to inform the HC's receptionist about s:

$$Bel_{Bill} Avail(HBS, Bill, HC, s) \wedge Bel_{Bill} Bel_{HBS} Goal_{Bill} Inf_{HBS,HC}(s).$$

Bill also believes that the HC's receptionist believes that HBS is a valid information source about s:

$$Bel_{Bill} Bel_{HC} Valid(HBS, HC, s).$$

Then, from theorem 6, we can infer that Bill trusts HBS to inform the HC's receptionist about s with regard to his goal that the HC's receptionist will believe s:

$$Trust(Bill, HBS, inf_{HC}(s), Bel_{HC} s).$$

The following theorem 7 highlights the relationship between trust in a communication system's inaction and the fact that the communication system ensures privacy. It says that: if i has the goal that z does not believe P, i believes that j ensures the privacy of information P from agent z, i believes that j believes that i wants j not to inform z about P and, i believes that z believes that j is a valid information source about P then, i trusts j not to inform z about P with regard to his goal that z does not believe P.

Theorem 7 *Let $i, j, z \in AGT$ and $inf_z(P) \in$ INFO, then:*
$\vdash (Goal_i X \neg Bel_z P \wedge Bel_i Priv(j, i, z, P) \wedge Bel_i Bel_j Goal_i \neg Inf_{j,z}(P) \wedge$
$Bel_i Bel_z Valid(j, z, P)) \rightarrow Trust(i, j, \neg inf_z(P), \neg Bel_z P)$

Example 7. In this version of the scenario of the Hotel Booking Service (HBS) Bill has the goal that a potential intruder will not have access to the information r ("Bill's credit card number is 01234567891234"):

$$Goal_{Bill} X \neg Bel_{intruder} r.$$

Bill believes that HBS ensures the privacy of information r from potential intruders and that HBS believes that Bill wants HBS not to inform a potential intruder about r:
$Bel_{Bill}Priv(HBS, Bill, intruder, r) \wedge Bel_{Bill}Bel_{HBS}Goal_{Bill}\neg Inf_{HBS,intruder}(r)$.
Finally, Bill believes that every potential intruder believes that HBS is a valid information source about credit card numbers:

$$Bel_{Bill}Bel_{intruder}Valid(HBS, intruder, r).$$

Then, from theorem 7, we can infer that Bill trusts HBS not to inform a potential intruder about r with regard to his goal that a potential intruder will not believe r:

$$Trust(Bill, HBS, \neg inf_{intruder}(r), \neg Bel_{intruder}r).$$

8 Related works

Several logical models of trust in information sources have been proposed in the recent literature [20, 18, 10, 9]. Some of them take the concept of trust as a primitive [20], whereas others reduce trust to a kind of belief of the truster [18, 10]. For instance, in [18] trust is characterized on the basis of two kinds of beliefs of the truster: the truster's belief that a certain rule or regularity applies to the trustee (called "rule belief"), and the truster's belief that the rule or regularity is going to be followed by the trustee (called "conformity belief"). Nevertheless, all these models ignore the motivational aspect of trust in information sources. Moreover, all these models do not consider the epistemic supports for this form of trust. In the present paper both aspects of trust in information sources have been taken into account. On the one hand, we have modeled the truster's epistemic goal of knowing whether a certain fact is true. On the other hand, we have modeled the properties of information sources such as sincerity, validity, competence, etc. and shown that some of them are epistemic supports for an agent's trust in an information source, that is, they are sufficient conditions for trusting an information source to inform whether a certain fact is true (theorem 4). As far as communication systems are concerned, there are several logical models in the literature which deal with the properties of a communication system such as privacy, confidentiality, availability, integrity, authentication (*e.g.*[8, 4]). Nevertheless, there is still no formal analysis of the relationships between these properties of a communication system and trust in the communication system. In this work such relationships have been clarified (theorems 6 and 7).

9 Conclusion

We have presented in a modal logical framework a model that integrates in the definition of trust: the truster's goal, the trustee's action that ensures the achievement of the truster's goal, and the trustee's ability and intention to do this action. In the same logical framework we have defined several properties of information sources (validity, completeness, sincerity, competence, vigilance and cooperativity) and discussed their relationships with an agent's trust in an information source. In the last part of the paper, we have investigated some properties of communication systems (availability and privacy) and discussed their relationships with an agent's trust in a communication system. It has to be noted that, due to the complexity of the concepts involved in our analysis of trust, we had to accept strong simplifications. For instance, in the definitions of the properties of information sources and communication systems entailment is formalized by a material implication, while some form of conditional might be more adequate. Our future work will be devoted to refine this aspect of the proposed logical formalization of trust.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.
2. P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, Cambridge, 2001.
3. M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, 1987.
4. M. Burrows, M. Abadi, and R. M. Needham. A logic of authentication. *ACM Transactions on Computer Systems*, 8(1):18–36, 1990.
5. C. Castelfranchi and R. Falcone. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proc. of the Third International Conference on Multiagent Systems (ICMAS'98)*, pages 72–79, 1998.
6. C. Castelfranchi and R. Falcone. Social trust: A cognitive approach. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer, 2001.
7. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
8. F. Cuppens and R. Demolombe. A Deontic Logic for Reasoning about Confidentiality. In *Proc. of Third International Workshop on Deontic Logic in Computer Science (DEON'96)*, pages 66–79. Springer, 1996.
9. M. Dastani, A. Herzig, J. Hulstijn, and L. van der Torre. Inferring trust. In *Proc. of Fifth Workshop on Computational Logic in Multi-agent Systems (CLIMA V)*, pages 144–160. Springer, 2004.
10. R. Demolombe. To trust information sources: a proposal for a modal logical framework. In C. Castelfranchi and Y-H. Tan, editor, *Trust and Deception in Virtual Societies*. Kluwer, 1999.
11. R. Demolombe. Reasoning about trust: a formal logical framework. In *Proc. of iTrust 2004*, pages 291–303. Springer, 2004.
12. R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.
13. M. Fasli. *Agent Technology for E-commerce*. Wiley & Sons, Chichester, 2007.
14. J. Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, University of Amsterdam, The Netherlands, 1999.
15. D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.
16. A. Herzig, E. Lorini, J. F. Hubner, J. Ben-Naim, C. Castelfranchi, R. Demolombe, D. Longin, and L. Vercouter. Prolegomena for a logic of trust and reputation. In *Proceedings of the Third International Workshop on Normative Multiagent Systems (NorMAS 2008)*, forthcoming.
17. T. G. Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agent and Multi-Agent Systems*, 13:119–154, 2006.
18. A. J. I. Jones. On the concept of trust. *Decision Support Systems*, 33(3):225–232, 2002.
19. C. M. Jonker and J. Treur. Formal analysis of models for the dynamics of trust based on experiences. In *Proceedings of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'99)*, pages 221–231. Springer, 1999.

20. C. J. Liau. Belief, information acquisition, and trust in multi-agent systems: a modal logic formulation. *Artificial Intelligence*, 149:31–60, 2003.
21. E. Lorini and R. Demolombe. Trust and norms in the context of computer security. In *Proceedings of the Ninth International Conference on Deontic Logic in Computer Science (DEON'08)*, forthcoming.
22. E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77.
23. S. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, Scotland, 1994.
24. S. A. McIlraith, T. C. Son, and H. Zeng. Semantic web services. *IEEE Intelligent Systems*, 16(2):46–53, 2001.
25. J.-J. C. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1-2):1–40, 1999.
26. J. Sabater and C. Sierra. Regret: a reputation model for gregarious societies. In *Proc. of AAMAS'02*, pages 475–482. ACM Press, 2001.
27. R. B. Scherl and H. Levesque. Knowledge, action, and the frame problem. *Artificial Intelligence*, 144:1–39, 2003.