

Trust and norms in the context of computer security: a logical formalization*

Emiliano Lorini, Robert Demolombe

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

Abstract. In this paper we present a logical model of trust in which trust is conceived as an expectation of the truster about some properties of the trustee. A general typology of trust is presented. We distinguish trust in the trustee's action from trust in the trustee's disposition (motivational or normative disposition); positive trust from negative trust. A part of the paper is devoted to the formalization of security properties and to the analysis of their relationships with trust.

1 Introduction

Techniques of computer security have been mainly designed in the perspective of protecting a computer system with respect to attacks of ill-intentioned users who want, for example, to access private data. To prevent these situations techniques have been developed, like cryptography, in order to reduce risks and to make that standard users trust the computer system. However, another kind of scenario may happen where the computer system has been designed to violate some regulations about privacy. For example, private data gathered for some applications may be sold to a company for advertising without users' authorization. In these kinds of scenario, even if the computer system guarantees that ill-intentioned users have no capability to violate the norms, standard users want to trust the computer system about the fact that it will not intentionally violate the norms. In this perspective, the issue is not to trust the effectiveness of computer science techniques (like cryptography) but to trust the fact that norms are not deliberately violated by the system. That was the initial motivation of the work presented in this paper.

Since trust is a complex mental attitude, the first step was to propose a clear definition in a logical framework which is presented in section 2. In section 3 we present a global view of trust in order to point out several refinements of this concept. In section 4 we focus on the trustee's intention to do, or not to do, a certain action for the truster. Then, in section 5, we refine this approach by analyzing the disposition of the trustee to perform a certain action for the truster which is called *willingness*. In section 6 computer security properties are defined and their normative dimension is discussed. That leads to define in section 7 the normative dispositions of the trustee toward the truster which are called *obedience* and *honesty*.

* This work is supported by the project "ForTrust: social trust analysis and formalization" funded by the french Agence Nationale de la Recherche (ANR).

2 A logic for trust reasoning

The logic \mathcal{L} we use to formalize the relevant concepts involved in our model of social trust is a multimodal logic which combines the expressiveness of a simple dynamic logic [13] with the expressiveness of a logic of mental attitudes [6, 21] and obligations [1, 3]. The syntactic primitives of the logic \mathcal{L} are the following:

- a nonempty finite set of agents $AGT = \{i, j, \dots\}$;
- a nonempty finite set of atomic actions $ACT = \{\alpha, \beta, \dots\}$;
- a set of atomic formulas $ATM = \{p, q, \dots\}$.

The language of \mathcal{L} is the set of formulas defined by the following BNF:

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

where p ranges over ATM , α ranges over ACT and i ranges over AGT .

The operators of our logic have the following intuitive meaning. $Bel_i\phi$: the agent i believes that ϕ ; $After_{i:\alpha}\phi$: after agent i does α , it is the case that ϕ ($After_{i:\alpha}\perp$ is read: agent i cannot do action α); $Does_{i:\alpha}\phi$: 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); $Goal_i\phi$: the agent i wants that ϕ holds; $Obg\phi$: it is obligatory that ϕ . The following abbreviations are given: $Can_i(\alpha) \stackrel{\text{def}}{=} \neg After_{i:\alpha}\perp$; $Int_i(\alpha) \stackrel{\text{def}}{=} Goal_i Does_{i:\alpha}\top$; $Perm\phi \stackrel{\text{def}}{=} \neg Obg\neg\phi$. We write $Can_i(\alpha)$ as an abbreviation of $\neg After_{i:\alpha}\perp$ in order to make explicit the fact that $\neg After_{i:\alpha}\perp$ stands for: agent i can do action α (i.e. i has the capacity/ability to do α). $Int_i(\alpha)$ stands for: the agent i intends to do α . $Perm\phi$ stands for: ϕ is permitted.

Models of our logic are tuples $M = \langle W, R, D, B, G, O, V \rangle$ where:

- W is a non empty set of possible worlds or states.
- R is a collection of binary relations $R_{i:\alpha}$ on W , one for every couple $i:\alpha$ where $i \in AGT$ and $\alpha \in ACT$. Given an arbitrary world $w \in W$, if $(w, w') \in R_{i:\alpha}$ then w' is a world which can be reached from world w through the occurrence of agent i 's action α .
- D is a collection of binary relations $D_{i:\alpha}$ on W , one for every couple $i:\alpha$ where $i \in AGT$ and $\alpha \in ACT$. Given an arbitrary world $w \in W$, if $(w, w') \in D_{i:\alpha}$ then w' is the *next* world of w which will be reached from w through the occurrence of agent i 's action α .
- B is a collection of binary relations B_i on W , one for every agent $i \in AGT$. Given an arbitrary world $w \in W$, if $(w, w') \in B_i$ then w' is a world which is compatible with agent i 's beliefs at world w .
- G is a collection of binary relations G_i on W , one for every agent $i \in AGT$. Given an arbitrary world $w \in W$, if $(w, w') \in G_i$ then w' is a world which is compatible with agent i 's goals at world w .
- O is a binary relation on W . Given an arbitrary world $w \in W$, if $(w, w') \in O$ then w' is a world which is ideal at world w .
- $V : ATM \rightarrow 2^W$ is a valuation function.

Truth conditions for atomic formulas, negation and disjunction are entirely standard. The following are truth conditions for the modal operators introduced before.

- $M, w \models \text{After}_{i:\alpha}\phi$ iff $M, w' \models \phi$ for all w' such that $(w, w') \in R_{i:\alpha}$.
- $M, w \models \text{Does}_{i:\alpha}\phi$ iff $\exists w'$ such that $(w, w') \in D_{i:\alpha}$ and $M, w' \models \phi$.
- $M, w \models \text{Bel}_i\phi$ iff $M, w' \models \phi$ for all w' such that $(w, w') \in B_i$.
- $M, w \models \text{Goal}_i\phi$ iff $M, w' \models \phi$ for all w' such that $(w, w') \in G_i$.
- $M, w \models \text{Oblig}\phi$ iff $M, w' \models \phi$ for all w' such that $(w, w') \in O$.

2.1 Properties of the operators

The operators Bel_i , $\text{After}_{i:\alpha}$, $\text{Does}_{i:\alpha}$, Goal_i and Oblig are supposed to be normal modal operators satisfying standard axioms and rules of inference of system K . Operators for belief of type Bel_i are supposed to be $KD45$ modal operators, whilst every operator for goal of type Goal_i is supposed to be a KD operator. Thus, we make assumptions about positive and negative introspection for beliefs and we suppose that beliefs and goals cannot be inconsistent. Operators for obligations of type Oblig are also supposed to be KD as in SDL (standard deontic logic) [1].¹

As far as actions are concerned, we assume that actions of the same agent and actions of different agents occur in parallel.

$$\mathbf{Alt}_{Act} \text{ Does}_{i:\alpha}\phi \rightarrow \neg \text{Does}_{j:\beta}\neg\phi$$

Axiom \mathbf{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 $\neg\phi$ will be true afterward.

We also suppose that the world is never static in our framework, that is, we suppose that always there exists some agent i and action α such that i is going to perform α .

$$\mathbf{Active} \bigvee_{i \in AGT, \alpha \in ACT} \text{ Does}_{i:\alpha} \top$$

Axiom \mathbf{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.²

$$X\phi \stackrel{\text{def}}{=} \bigvee_{i \in AGT, \alpha \in ACT} \text{ Does}_{i:\alpha}\phi$$

The following Axiom \mathbf{Inc}_{Act} relates the operator $\text{Does}_{i:\alpha}$ with the operator $\text{After}_{i:\alpha}$.

$$\mathbf{Inc}_{Act, PAct} \text{ Does}_{i:\alpha}\phi \rightarrow \neg \text{After}_{i:\alpha}\neg\phi$$

According to $\mathbf{Inc}_{Act, PAct}$, if i is going to do α and ϕ will be true afterward, then it is not the case that $\neg\phi$ is true after i does α .

The following axioms relating intentions with actions seem quite natural in the case of intentional actions.

$$\begin{aligned} \mathbf{IntAct1} \quad & (\text{Int}_i(\alpha) \wedge \text{Can}_i(\alpha)) \rightarrow \text{Does}_{i:\alpha} \top \\ \mathbf{IntAct2} \quad & \text{Does}_{i:\alpha} \top \rightarrow \text{Int}_i(\alpha) \end{aligned}$$

¹ Semantic constraints corresponding to the axioms presented in this section are given in [12].

² Note that X satisfies the standard property $X\phi \leftrightarrow \neg X\neg\phi$.

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 [20, 19] in which a logical model of the relationships between intention and action performance is proposed.

As far as beliefs and goals are concerned, we only suppose that the two kinds of mental attitudes must be compatible, that is, if an agent has the goal that ϕ he cannot believe that $\neg\phi$. Indeed, the notion of goal we characterize is a notion of an agent's *chosen goal*, i.e. a goal that an agent decides to pursue. As some authors have stressed [2], a rational agent cannot decide to pursue a certain state of affairs ϕ , if he believes that $\neg\phi$ (this is called *weak realism* hypothesis).

$$\mathbf{WR} \quad Goal_i\phi \rightarrow \neg Bel_i\neg\phi$$

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

$$\mathbf{PIintr} \quad Goal_i\phi \rightarrow Bel_iGoal_i\phi$$

$$\mathbf{NIintr} \quad \neg Goal_i\phi \rightarrow Bel_i\neg Goal_i\phi$$

The following axiom relates obligations with beliefs:

$$\mathbf{BelObg} \quad Obg\phi \rightarrow Bel_iObg\phi$$

This axiom is based on the assumption that every agent has complete information of what is obligatory. It is justified by the fact that if it is expected that an agent does every action which is obligatory, he must have a complete information about what is obligatory. Note that by Axiom **BelObg**, the definition of the permission operator *Perm* and Axiom *D* for Bel_i , the following formula can be derived as a consequence: $Bel_iPerm\phi \rightarrow Perm\phi$. This means that in our logical framework every agent has sound information of what is permitted.

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} .

3 A global view of trust

In the present logical model trust is conceived as a complex configuration of mental states in which there is a main and primary motivational component (the principal reason activating the truster's delegating behavior): the goal to achieve some state of affairs φ (the trust in the trustee is always relative to some interest, need, concern, desire of the truster); and a complex configuration of truster's beliefs about the qualities of the trustee. On this point we agree with Castelfranchi & Falcone [4, 5] on the fact that a model of social trust must account for the truster's attribution process, that is, it must account for the truster's ascription of specific properties to the trustee (abilities, willingness, dispositions, etc.) and the truster's ascription of properties to the environment in which the trustee is going to act (will the environmental conditions prevent the trustee from accomplishing the task that the truster has delegated to him?). From this perspective there is a pressing need for elaborating richer models of social trust in which the

truster's expectation and its components are explicitly modeled. To this end, we present in the following sections a conceptual and logical model of social trust which shows that trust is not a unitary and simplistic notion. More precisely, we assume that i 's trust in agent j necessarily involves a main and primary motivational component which is a goal of the truster. If i trusts agent j then necessarily i trusts j with respect to some of his goals. Moreover, the core of trust is a belief of the truster about some properties of the trustee, that is, if 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 aim of the following sections is to clarify the nature of such a belief of the truster.

| | Trust about action | Trust about disposition | |
|-----------------|-----------------------------------|---|--|
| | | Motivational | Normative |
| Positive | i trusts j to do α | i trusts j to be willing to do α for him | i trusts j to be obedient to do α |
| Negative | i trusts j not to do α | i trusts j to be willing not to do α for him | i trusts j to be honest to do α |

Table 1. Typology of Trust.

We also claim that there is no unique definition of trust, but there are several types of trust depending on the kinds of properties that the truster ascribes to the trustee. The ontology of trust proposed in the following sections is organized according to two main dimensions (see Table 1). First, we distinguish between *positive trust* and *negative trust*. In positive trust i is focused on the domain of gains (goal achievements) whereas in negative trust i is focused on the domain of losses (goal frustrations). The second distinction is between *trust in the trustee's actions* and *trust in the trustee's dispositions*. In the former case, i 's trust in j is based on i 's belief that j will perform (resp. refrain from performing) a certain action α ; whereas in the latter case i 's trust in j is based on i 's belief that j is disposed to perform (resp. to refrain from performing) a certain action α . By combining the previous two dimensions we characterize four general categories of trust.

- i trusts j because i believes that j can help him to achieve a certain goal by performing a certain action α and j is going to perform action α (*i 's positive trust in j 's action*);
- i trusts j because i believes that j is in the condition to damage him (i.e. to frustrate a goal of i) by doing a certain action α and j will refrain from performing action α (*i 's negative trust in j 's action*);
- i trusts j because i believes that j can help him to achieve a certain goal by performing a certain action α and j is disposed to perform action α (*i 's positive trust in j 's disposition*);
- i trusts j because i believes that j is in the condition to damage him (i.e. to frustrate a goal of i) by doing a certain action α and j is disposed to refrain from performing action α (*i 's negative trust in j 's disposition*).

We introduce a further sophistication by distinguishing between motivational dispositions and normative (or moral) dispositions of the trustee. Indeed, in the context of i 's positive trust in j 's disposition (resp. i 's negative trust in j 's disposition), j 's disposition to perform a certain action α (resp. j 's disposition to refrain from performing a certain action α), can be interpreted in two different ways. According to the motivational interpretation, i 's belief that j is disposed to perform action α (resp. j is disposed to refrain from performing action α) stands for i 's belief that j is willing to do action α for him (resp. j is willing not to do action α for him). According to the normative interpretation, i 's belief that j is disposed to perform action α (resp. j is disposed to refrain from performing action α) stands for i 's belief that j will obey to the obligation of doing action α (resp. will not perform action α if he has no permission to perform action α). Thus, our ontology of trust gets refined in such a way that we can distinguish two different types of i 's positive trust in j 's disposition and two different types of i 's negative trust in j 's disposition. Namely: *i 's positive trust in j 's motivational disposition*, *i 's negative trust in j 's motivational disposition*, *i 's positive trust in j 's moral disposition*, *i 's negative trust in j 's moral disposition*.

The concepts of positive and negative trust in the trustee's action are studied in section section 4. Section 5 is devoted to the analysis of positive and negative trust in the trustee's motivational disposition. The reader must wait until section 7 for positive and negative trust in the trustee's normative disposition.

3.1 Some related works

Our logical model of trust shares some intuitions with Castelfranchi & Falcone's conceptual and informal model of trust [4, 5]. As emphasized in the previous section, 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 sustained by the repeated direct interactions with other agents under the assumption that iterated experiences of success strengthen the trustor's confidence [17]. More sophisticated models of social trust have been developed in which reputational information is added to information obtained via direct interaction (e.g. [14]). All these trust models are in our view over-simplified since they do not consider the indirect supports for the trust expectation. Trust is rather a complex expectation of the trustor about some properties of trustee which are relevant for the achievement of goal of the trustor.

Nevertheless, there are important difference between our model of trust and Castelfranchi & Falcone's model. For instance, we think that their model of trust is not sufficiently clear in distinguishing trust in the trustee's actions and trust in the trustee's willingness. This distinction is for us fundamental since it allows to capture two forms of trust which have different natures. Moreover, their model only account for positive trust and do not consider negative trust.

As far as logics of trust are concerned, we think that there is still no comprehensive logical model of this social phenomenon. Indeed, logical models of trust have been focused almost exclusively on trust in information sources (informational trust) [18, 16, 10, 8], or they have reduced trust to a certain kind of beliefs neglecting the motivational aspects of trust [9]. In [9] trust is defined as a trustor's sort of belief, called "strong

belief”, about some properties of the trustee. They may be epistemic properties, like sincerity or competence, dynamic properties, like ability, or deontic properties like obedience and honesty. From this perspective there is a pressing need for elaborating more general logical models of social trust in which the truster’s expectation and its different components are explicitly modeled, and in which the motivational aspect of trust is taken into account.

4 Trust in the trustee’s action

We first define the notion of positive trust in the trustee’s action. Such a notion presents four different arguments: truster, trustee, truster’s goal, trustee’s action.

Definition 1 POSITIVE TRUST ABOUT 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:*³

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*

Condition 1 concerns the trustee’s power to satisfy the truster’s goal that ϕ by means of the performance of action α . Conditions 2 and 3 are about the trustee’s properties which are necessary and sufficient for him to perform action α . The formal translation of Definition 1 is:

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

In our logic the second and third condition in the definition of positive trust are together equivalent to $Does_{j:\alpha}\top$ (by Axiom **IntAct2**), so the definition of trust can be simplified as follows:

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

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

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 1 *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

$$\vdash ATrust(i, j, \alpha, \phi) \rightarrow Bel_i X\phi$$

The dual notion of negative trust in the trustee’s action 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 . That leads to the following definition.

³ 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.

Definition 2 NEGATIVE TRUST ABOUT ACTION. *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 $\neg\phi$ AND*
2. *j has the capacity to do α AND*
3. *j does not intend to do α*

The formal translation of definition 2 is given by the following abbreviation.

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

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

5 Trust in the trustee's disposition: the motivational case

The fact that agent *j* intends to do α may be a consequence of his willingness with regard to *i*'s intention that *j* does α . That leads to define the more specific notions of positive and negative trust in the trustee's willingness. Indeed, *i*'s trust in *j* does not necessarily depend on *i*'s ascription of an actual intention to *j* to do a certain action α . There are forms of trust which are based on *i*'s ascription of a potential intention to *j*. In these cases *i* attributes to *j* a positive disposition which is called *j*'s *willingness*. More precisely, we suppose that *j* is *willing to do the action α for i* if and only if *j* has the conditional goal (or conditional intention) to form the intention to perform action α under the condition in which he believes that *i* wants him to do α . Thus, *willingness* is interpreted here as closely related to the concept of *goal adoption*. In this perspective, saying "*j* is willing to do everything for *i*" means "*j* wants to do whatever *i* wants him to do" and saying "*j* is willing to do action α for *i*" means "*j* wants to do α in case *i* wants him to do α ".⁴ The following abbreviation captures our notion of willingness in a formal way.

$$Will_{j,i}(\alpha) \stackrel{\text{def}}{=} Goal_j(Bel_j Goal_i Does_{j:\alpha} \top \rightarrow Int_j(\alpha)) \wedge \\ \neg Goal_j \neg Bel_j Goal_i Does_{j:\alpha} \top$$

where $Will_{j,i}(\alpha)$ stands for: *j is willing to do α for i*. The second condition in the definition of willingness is given in order to prevent from saying that *j* is willing to do α for *i*, when *j* wants not to believe that *i* does not want him to do α .

We define a related concept of *j*'s willingness not to do α for *i*. According to our definition, *j is willing not to do the action α for i* if and only if *j* has the conditional goal that he will not have the intention to do action α unless he believes that *i* does not want him not to do α .

$$Will_{j,i}(\neg\alpha) \stackrel{\text{def}}{=} Goal_j(Int_j(\alpha) \rightarrow Bel_j \neg Goal_i \neg Does_{j:\alpha} \top) \wedge$$

⁴ *Willingness* may have different natures. Agent *i* might be willing to do a certain action α for *j* since he expects that if he does α , he will get something in return by *j*; or *i* might be willing to do a certain action α for *j* since he expects that if he does not do α , *j* will do something bad for him, etc. In this work we focus on the core of the concept of willingness without investigating the more specific forms of willingness (i.e. the reasons to be willing).

$$\neg Goal_j Bel_j \neg Goal_i \neg Does_{j:\alpha} \top$$

where $Will_{j,i}(\neg\alpha)$ stands for: j is willing not to do α for i .⁵ The following two theorems highlight some interesting properties of our concept of willingness.

Theorem 2 *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

1. $\vdash Will_{j,i}(\alpha) \rightarrow (Bel_j Goal_i Does_{j:\alpha} \top \rightarrow Int_j(\alpha))$
2. $\vdash Will_{j,i}(\neg\alpha) \rightarrow (Int_j(\alpha) \rightarrow Bel_j \neg Goal_i \neg Does_{j:\alpha} \top)$

According to Theorem 2.1, if j is willing to do α for i and j believes that i wants him to do α , then j will adopt i 's goal in such a way that he will intend to do α . In this sense Theorem 2.1 captures the *adoptive* process which leads from a j 's positive disposition toward i to the situation in which j intends to do what i wants him to do. According to Theorem 2.2, if j is willing not to do α for i and intends to do action α , then he has to believe that i does not want him not to do α .

From the the concept of willingness, we can characterize the concept of i 's positive trust in j 's willingness.

Definition 3 POSITIVE TRUST ABOUT WILLINGNESS. *i trusts j about j 's willingness to do α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:*

1. j , by doing α , will ensure that ϕ AND
2. j has the capacity to do α AND
3. j is willing to do α for i

Formally:

$$WTrust(i, j, \alpha, \phi) \stackrel{\text{def}}{=} Goal_i X\phi \wedge Bel_i (After_{j:\alpha} \phi \wedge Can_j(\alpha) \wedge Will_{j,i}(\alpha))$$

where $WTrust(i, j, \alpha, \phi)$ stands for: i trusts j about j 's willingness to do α with regard to his goal that ϕ . The following theorem highlights the relationship between the notions of $ATrust(i, j, \alpha, \phi)$ and $WTrust(i, j, \alpha, \phi)$.

Theorem 3 *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

1. $\vdash (Bel_i Bel_j Goal_i Does_{j:\alpha} \top \wedge WTrust(i, j, \alpha, \phi)) \rightarrow Bel_i Int_j(\alpha)$
2. $\vdash (Bel_i Bel_j Goal_i Does_{j:\alpha} \top \wedge WTrust(i, j, \alpha, \phi)) \rightarrow ATrust(i, j, \alpha, \phi)$

For instance, according to Theorem 3.2, if i trusts j about j 's willingness to do α with regard to his goal that ϕ and i believes that j believes that i wants j to do α , then i trusts j to do α with regard to his goal that ϕ .

The concept of negative trust in the trustee's willingness can be defined as follows.

⁵ As for the definition of j 's willingness to do α for i , we add the condition $\neg Goal_j Bel_j \neg Goal_i \neg Does_{j:\alpha} \top$ in order to prevent from saying that j is willing not do α for i , when j wants to believe that i does not want that he does not do action α . The same solution is adopted in section 7 for the definitions of obedience and honesty.

Definition 4 NEGATIVE TRUST ABOUT WILLINGNESS. *i trusts j about j's willingness not to do α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:*

1. *j, by doing α , will ensure that $\neg\phi$ AND*
2. *j has the capacity to do α AND*
3. *j is willing not to do α for i*

Formally:

$$WTrust(i, j, \neg\alpha, \phi) \stackrel{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\neg\phi \wedge Can_j(\alpha) \wedge Will_{j,i}(\neg\alpha))$$

where $WTrust(i, j, \alpha, \phi)$ stands for: *i trusts j about j's willingness not to do α with regard to his goal that ϕ .*

The following theorem highlights the relationship between negative trust about willingness and negative trust about action. It says that: negative trust about willingness entails negative trust about action in the context where *i* believes that *j* does not believe that *i* does not want *j* not to do α .

Theorem 4 *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

$$\vdash (Bel_i \neg Bel_j \neg Goal_i \neg Does_{j:\alpha} \top \wedge WTrust(i, j, \neg\alpha, \phi)) \rightarrow ATrust(i, j, \neg\alpha, \phi)$$

6 Norms in computer security

In the field of computer science the notion of security may have two different meanings: there is no computer failure, or there is no violation of norms about computer usage. In this paper we adopt the second meaning. Here agents may be human agents or software agents. In the case of software agents, we talk about their mental attitudes like beliefs or intentions and we assume that their actions are intentional actions. Moreover, we suppose that for a software agent, performing an action means executing a program, and a certain program is performed by the software agent only if the effects of its execution conform to what has been specified by the designer of the program. For instance, a software agent can inform someone about something only by performing the act *inform* which is the procedure specified by the designer as a means for inducing someone to believe something. It cannot inform someone about something by performing some sequence of *insert* actions or *delete* actions since this is not the procedure specified by the designer.

In this work the security properties that should be guaranteed are restricted to: integrity, availability and privacy [7]. For simplification, we have ignored properties like: authentication or non repudiation. As a matter of simplification we have only considered computer systems of the kind information systems (for instance a database system). A similar analysis could be done for transmission systems (for instance Internet).

In order to study security properties we extend the logic \mathcal{L} with the following specific actions: $inf_j(\phi)$ (action of informing *j* about ϕ), $ins_j(\phi)$ (action of inserting the information ϕ in *j*), $del_j(\phi)$ (action of deleting the information ϕ from *j*), $ask_j(\alpha)$ (action of asking *j* to do action α). The following abbreviations are given

for denoting the performance of the previous special actions by an arbitrary agent i : $Inf_{i,j}(\phi) \stackrel{\text{def}}{=} Does_{i:inf_j(\phi)}\top$; $Ins_{i,j}(\phi) \stackrel{\text{def}}{=} Does_{i:ins_j(\phi)}\top$; $Del_{i,j}(\phi) \stackrel{\text{def}}{=} Does_{i:del_j(\phi)}\top$; $Ask_{i,j}(\alpha) \stackrel{\text{def}}{=} Does_{i:ask_j(\alpha)}\top$.

The constructions $Inf_{j,i}(\phi)$, $Ins_{i,j}(\phi)$, $Del_{i,j}(\phi)$ are used to describe the interaction between an information system j and an agent i (i may be a human agent or a software agent). $Inf_{j,i}(\phi)$ means: the information system j informs agent i about ϕ . $Ins_{i,j}(\phi)$ means: agent i inserts the information ϕ in the information system j (or i makes that j believes that ϕ). $Del_{i,j}(\phi)$ means: agent i deletes the information ϕ from the information system j (or i makes that j does not believe that ϕ). For human agents or software agents the construction $Ask_{i,j}(\alpha)$ expresses that: agent i asks j to do the action α . In the following sections security properties are going to be defined.

6.1 Security properties

Definition 5 *The information system j guarantees the privacy of information ϕ if and only if for every agent k , if j informs k about ϕ , then it is permitted that j informs k about ϕ .*

Formally,

$$Priv_j(\phi) \stackrel{\text{def}}{=} \bigwedge_{k \in AGT} (Inf_{j,k}(\phi) \rightarrow PermInf_{j,k}(\phi))$$

where $Priv_j(\phi)$ stands for: the information system j guarantees the privacy of information ϕ .

Definition 6 *The information system j guarantees the integrity of information ϕ if and only if for every agent k , if k inserts (resp. deletes) ϕ , then it is permitted that k inserts (resp. deletes) ϕ .*

Formally,

$$Intg_j(\phi) \stackrel{\text{def}}{=} \bigwedge_{k \in AGT} (Ins_{k,j}(\phi) \rightarrow PermIns_{k,j}(\phi)) \wedge \bigwedge_{k \in AGT} (Del_{k,j}(\phi) \rightarrow PermDel_{k,j}(\phi))$$

where $Intg_j(\phi)$ stands for: the information system j guarantees the integrity of information ϕ .

Definition 7 *Agent i guarantees the availability to do the action α for j if and only if, if i has the right to oblige j to do α and i asks j to do α , then j does α .*

Formally,

$$Avail_{i,j}(\alpha) \stackrel{\text{def}}{=} (Right_{i,j}(\alpha) \wedge Ask_{i,j}(\alpha)) \rightarrow Does_{j:\alpha}\top$$

where $Avail_{i,j}(\alpha)$ stands for: agent i guarantees the availability to do the action α for j , and

$$Right_{i,j}(\alpha) \stackrel{\text{def}}{=} Ask_{i,j}(\alpha) \rightarrow ObgDoes_{j:\alpha}\top$$

The intuitive meaning of $Right_{i,j}(\alpha)$ is that by asking j to do α i “creates” the obligation for j to do α .

7 Trust in the trustee's disposition: the normative case

In the context of computer security the fact that agent j intends to do α may be a consequence of his fulfillment of the obligation to do this action. In this case we say that j is *obedient*. In a similar way, the fact that he does not intend to do α may be a consequence of the fact that he respects the prohibition to do this action. In this case we say that j is *honest*. It is worth noting that there is a deep analogy between the fact that i 's goal is that j does α (resp. it is not the case that i 's goal is that j does not do α) and the fact that it is obligatory that j does α (resp. it is permitted that j does α). The justification of this analogy is that what is obligatory can be interpreted as the goal of people who institute the norms, and what is permitted as what is possible with respect to their goal. In the formal definitions below, this analogy is expressed by the fact that the definition of obedience (resp. honesty) can be obtained from the definition of willingness to do (resp. willingness not to do) given in section 5 by substituting $ObgDoes_{j:\alpha}\top$ (resp. $PermDoes_{j:\alpha}\top$) to $Goal_i Does_{j:\alpha}\top$ (resp. $\neg Goal_i \neg Does_{j:\alpha}\top$). In the following this analogy will be called “*motivational / normative analogy*”.

On the one hand we suppose that j is *obedient to do the action* α if and only if, j has the conditional goal that if he believes that it is obligatory that he does α , then he intends to do α . Formally,

$$Obed_j(\alpha) \stackrel{\text{def}}{=} Goal_j(Bel_j ObgDoes_{j:\alpha}\top \rightarrow Int_j(\alpha)) \wedge \\ \neg Goal_j \neg Bel_j ObgDoes_{j:\alpha}\top$$

where $Obed_j(\alpha)$ stands for: j is obedient with regard to the obligation to do the action α .

On the other hand we suppose that j is *honest to do the action* α if and only if, j has the conditional goal that if he has the intention to do α , then he believes that it is permitted that he does α . Formally,

$$Honst_j(\alpha) \stackrel{\text{def}}{=} Goal_j(Int_j(\alpha) \rightarrow Bel_j PermDoes_{j:\alpha}\top) \wedge \\ \neg Goal_j Bel_j PermDoes_{j:\alpha}\top$$

where $Honst_j(\alpha)$ stands for: j is honest with regard to the permission to do the action α .

The following two theorems highlight some interesting properties of the concepts of obedience and honesty.

Theorem 5 *Let $j \in AGT$ and $\alpha \in ACT$. Then:*

1. $\vdash Obed_j(\alpha) \rightarrow (Bel_j ObgDoes_{j:\alpha}\top \rightarrow Int_j(\alpha))$
2. $\vdash Honst_j(\alpha) \rightarrow (Int_j(\alpha) \rightarrow Bel_j PermDoes_{j:\alpha}\top)$

According to Theorem 5.1, if j is obedient with regard to the obligation to do the action α and believes that it is obligatory to do α , then j will adopt such an obligation in such a way that he will intend to do α . Theorem 5.1, which is symmetrical to Theorem 2.1 for willingness captures the *adoptive* process which leads from j 's obedience to the

situation in which j intends to do what is obligatory to do. According to Theorem 5.2 (which is symmetrical to Theorem 2.2 for willingness), if j is honest with regard to the permission to do the action α and intends to do action α , then he has to believe that it is permitted to do action α .

We are now in the position to define a concept of i 's trust in j 's obedience which is symmetrical to the concept of i 's positive trust in j 's willingness given in section 5.

Definition 8 TRUST ABOUT OBEDIENCE. i trusts j to be obedient in doing α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:

1. j , by doing α , will ensure that ϕ AND
2. j has the capacity to do α AND
3. j is obedient in doing α

Formally,

$$OTrust(i, j, \alpha, \phi) \stackrel{\text{def}}{=} Goal_i X\phi \wedge Bel_i(After_{j:\alpha}\phi \wedge Can_j(\alpha) \wedge Obed_j(\alpha))$$

where $OTrust(i, j, \alpha, \phi)$ stands for: i trusts j to be obedient in doing α with regard to his goal that ϕ . The following theorems highlight the relationships between trust about obedience and positive trust about action, and between trust about obedience and the property of availability.

Theorem 6 Let $i, j \in AGT$ and $\alpha \in ACT$. Then:

1. $\vdash (Bel_i(Right_{i,j}(\alpha) \wedge Ask_{i,j}(\alpha)) \wedge OTrust(i, j, \alpha, \phi)) \rightarrow ATrust(i, j, \alpha, \phi)$
2. $\vdash OTrust(i, j, \alpha, \phi) \rightarrow Bel_i Avail_{i,j}(\alpha)$

The intuitive meaning of Theorem 6.1 is that trust about obedience entails positive trust about action in the context where i believes that he has the right to oblige j to do α and he exercises his right. Theorem 6.2 means that trust about obedience entails that i believes that the availability to do α is guaranteed by j . Notice that in this theorem i 's goal is not $Avail_{i,j}(\alpha)$. The goal ϕ may be any situation which can be obtained by doing α . For instance, i 's goal may be to know meteorological forecasts and the action α is that j informs i about these expectations. Then, in that example, the theorem 6.2 says that the consequence of i 's trust in j 's obedience to do α is that i believes that j guarantees the availability to inform him about meteorological forecasts.

We now define a concept of i 's trust in j 's honesty which is symmetrical to the concept of i 's negative trust in j 's willingness given in section 5.

Definition 9 TRUST ABOUT HONESTY. i trusts j to be honest in doing α with regard to his goal that ϕ if and only if i wants ϕ to be true and i believes that:

1. j , by doing α , will ensure that $\neg\phi$ AND
2. j has the capacity to do α AND
3. j is honest in doing α

Formally,

$$HTrust(i, j, \alpha, \phi) \stackrel{\text{def}}{=} Goal_i X \phi \wedge Bel_i(After_{j:\alpha} \neg \phi \wedge Can_j(\alpha) \wedge Honst_j(\alpha))$$

where $HTrust(i, j, \alpha, \phi)$ stands for: i trusts j to be honest in doing α with regard to his goal that ϕ .

We denote with $IAct(\psi)$ the set of all actions of informing some agent about ψ . In formal terms: $IAct(\psi) \stackrel{\text{def}}{=} \{inf_z(\psi) : z \in AGT\}$. Then, the following two theorems can be derived.

Theorem 7 *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

1. $\vdash (Bel_i \neg Bel_j PermDoes_{j:\alpha} \top \wedge HTrust(i, j, \alpha, \phi)) \rightarrow ATrust(i, j, \neg \alpha, \phi)$
2. $\vdash \bigwedge_{\alpha \in IAct(\psi)} (HTrust(i, j, \alpha, \phi)) \rightarrow Bel_i Priv_j(\psi)$

Theorem 7.1, which is symmetrical to Theorem 4 for negative trust about willingness, means that trust about honesty entails negative trust about action in the context where i believes that j does not believe that he has the permission to do α . Theorem 7.2 means that i 's trust in j 's honesty for every action of informing an agent about ψ entails that i believes that the privacy for ψ is guaranteed by j . Like in Theorem 6.2, in Theorem 7.2 i 's goal is not $Priv_j(\psi)$. A theorem similar to Theorem 7.2 can be proved for the property of integrity of an information system j ($Intg_j(\phi)$) since the set of permitted actions is explicitly defined.

8 Conclusion

The logical framework which has been presented allows to give precise definitions to several sophisticated notions of trust, going from a general one to more specific ones which are relevant to the context of computer security. In addition, theorems have been proved which give sufficient conditions about obedience or honesty to guarantee that an agent can believe that security properties hold. The benefits of the logical formalization are manifold. It points out some facts that may look as trivialities but that may be left implicit without the help of this formal framework. For instance, consequences that an agent can infer from what he trusts are just beliefs not truth. That is inherent to the notion of trust. Also, it raises some non trivial questions.

Due to the complexity of the involved concepts we had to accept strong simplifications. The first one is that our formal definition of the concept of obligation is very crude. The second is that in some definitions entailment is formalized by a material implication in the scope of goal modalities, while some form of conditional might be more adequate. The same comment applies to the definition of right where a "counts as" conditional [15] would be more appropriate than material implication. Also, security properties have been defined for a specific proposition, while these properties are usually expected for a set of proposition about a given topic, and a more realistic notion of trust should be based on several degrees of trust. Finally, we almost ignored the temporal dimension. In many cases trust is about a trustee's property which is not contingent to the current situation, but holds for some period of time. All these issues require future investigations, but we believe that to analyze so complex problems it was better to start with simple assumptions, even if they can be seen as oversimplifications.

References

1. L. Åqvist. Deontic logic. In D. M. Gabbay and F. Geunther, editors, *Handbook of Philosophical Logic*. Kluwer Academic Publishers, Dordrecht, 2002.
2. M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, 1987.
3. J. Carmo and A. Jones. Deontic Logic and Contrary-to-Duties. In D. Gabbay, editor, *Handbook of Philosophical Logic (Rev. Edition)*. Reidel, to appear.
4. 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.
5. 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.
6. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
7. 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 72–79, 1996.
8. 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)*, volume LNCS 3487, pages 144–160. Springer, 2004.
9. R. Demolombe. Reasoning about trust: a formal logical framework. In *Trust management: Second International Conference iTrust*, volume LNCS 2995, pages 291–303. Springer.
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*, pages 111–124. Kluwer, 2001.
11. R. Demolombe and C.-J. Liau. A logic of graded trust and belief fusion. In *Proc. of Fourth Workshop on Deception, Fraud and Trust in Agent Societies*, pages 13–25, 2001.
12. R. Demolombe and E. Lorini. A logical account of trust in information sources. In *Proc. of the Eleventh International Workshop on Trust in Agent Societies*, to appear.
13. D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.
14. 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.
15. A. J. Jones and M. Sergot. A formal characterisation of institutionalised power. *Journal of the Interest Group in Pure and Applied Logics*, 4(3), 1996.
16. A. J. I. Jones and B. S. Firozabadi. On the characterization of a trusting agent: Aspects of a formal approach. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer, 2001.
17. C. M. Jonker and J. Treur. Formal analysis of models for the dynamics of trust based on experiences. In *Proc. of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'99)*, volume 1647 of LNCS, pages 221–231. Springer, 1999.
18. C. J. Liau. Belief, information acquisition, and trust in multi-agent systems: a modal logic formulation. *Artificial Intelligence*, 149:31–60, 2003.
19. E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, to appear.
20. E. Lorini, A. Herzig, and C. Castelfranchi. Introducing “attempt” in a modal logic of intentional action. In *Logics in Artificial Intelligence: 10th European Conference (JELIA 2006)*, volume LNAI 4160, pages 280–292. Springer, 2006.
21. 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.