Trust within the Context of Organizations: A Formal Approach

Emiliano Lorini, Rino Falcone, Cristiano Castelfranchi

Trust within the Context of Organizations: A Formal Approach

Emiliano Lorini[1], Rino Falcone[2], and Cristiano Castelfranchi[2]

[1] Institut de Recherche en Informatique de Toulouse (IRIT), France
[2] Institute of Cognitive Sciences and Technologies-CNR, Italy
Emiliano.Lorini@irit.fr,
rino.falcone@istc.cnr.it,
cristiano.castelfranchi@istc.cnr.it

Abstract. We present in this paper a logical model of trust within organizations. Three forms of trust are investigated: trust in an agent (i.e. interpersonal trust), trust in a role, trust in an agent *qua* player of a role. The relationships between the three forms of trust are investigated. A part of the paper is devoted to the analysis of trust of an authority (e.g. an employer) in a subordinate (e.g. an employee).

1 Introduction

When looking at human organizations, social scientists have been mostly interested in individuating the antecedents of collective behavior and collective action between interacting individuals and roles. A central concern of the field has been identifying the determinants of intraorganizational cooperation, coordination and delegation [20,2]. Among the different determinants, trust has been recognized as one of the most important [7,16].

In this paper, we will study trust and organizations from the perspective of computer scientists working in the field of multi-agent systems (MAS). Indeed, to provide a formal analysis of trust within the context of organizations is of definite importance for the theory and development of multi-agent systems. In the recent years, in the MAS field there has been a growing interest in the theory of organization. Several formal approaches to the characterization of organizational concepts have been proposed [22,10] as well as general methodologies for MAS [23,13] which are based on organizational concepts as their cornerstones and which provide the guidelines for the specification and the design of MAS environments. In these formal approaches and existing methodologies, a multi-agent system is conceived as an organization consisting of various interacting roles which can be played by different agents. Although the concept of organization has been extensively studied in the agent domain, there is still no comprehensive formal account of the issue of trust in agent organizations. For instance, the distinction between the concept of *trust in an agent* and the concept of *trust in a role* is not clearly and deeply analyzed. Indeed, most of formal models of trust proposed in the agent domain have a limited perspective and only focus on trust in information sources in the specific context of information exchange between agents (e.g. [17,14,8]). The aim of the present paper is to extend our conceptual and formal model of social trust [18,9] to

the analysis of trust within organizations. This is in order to fill an existing gap in the literature about formal models of agents and multi-agent systems.

In particular, we will present in this paper a logical model of trust within the context of organizations. We model organizations as social entities in which agents play roles. Individual agents are described in terms of their mental attitudes (beliefs, goals, intentions). In an organization there are different roles to which certain powers are assigned. When an agent plays a certain role, he inherits the powers assigned to the role. We study trust at three different levels of generality. We start with the more general concept of an agent i's trust in another agent j abstracting away from the concept of role (*interpersonal trust*). We conceive interpersonal trust as an agent's disposition which is reducible to his beliefs and goals. In particular, we define trust in terms of a goal of the truster and the truster's belief that the trustee has the right properties (powers, abilities, dispositions) to ensure that his goal will be achieved. Then, we introduce the concept of role in order to investigate what it means that *an agent i trusts a certain role x* and *an agent i trusts another agent j qua player of a certain role x*. We focus on the relationships between the three different forms of trust (interpersonal trust, trust in a role and trust in an agent *qua* player of a role).

The paper is organized as follows. We start in Section 2 with a presentation of a modal logic which enables reasoning about actions and mental attitudes of agents (beliefs, goals and intentions), and about the roles that the agents play within the context of the organization. This logic will be used during the paper for formalizing the relevant concepts of our model of trust. The second part of the paper (Section 3) is devoted to present the three general concepts of trust that are relevant for a theory of organizations and for modeling and designing artificial organizations of agents: interpersonal trust (Section 3.1), trust in a role and and trust in an agent *qua* player of a role (Section 3.2). In Section 3.3, the three concepts are applied to the specific case of trust of an authority (e.g. an employer) in a subordinate (e.g. an employee). We conclude with a discussion of some directions for future works.

2 A Modal Logic of Mental Attitudes, Actions and Roles

We present in this section the multimodal logic $\mathcal{L}$ that we use to formalize the relevant concepts of our model of trust. $\mathcal{L}$ combines the expressiveness of dynamic logic [11] with the expressiveness of a logic of agents' mental attitudes [6]. Moreover, it enables reasoning about the relationships between different roles in the organization.

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 set of atomic formulas $ATM = \{p, q, \ldots\}$;
- a finite set of social roles $ROLE = \{x, y, \ldots\}$.

We add two additional formal constructions in order to specify the relationships between agents and roles and among different roles.

– a function $\mathscr{F}_{play} : ROLE \longrightarrow 2^{AGT} \setminus \emptyset$ which maps every role to a non-empty set of agents;
– a function $\mathscr{F}_{control} : ROLE \times ROLE \longrightarrow 2^{AT}$ which maps every couple of roles to a set of atomic actions.

Given a role $x \in ROLE$ and a non-empty set of agents $C \in 2^{AGT}$, $\mathscr{F}_{play}(x) = C$ means that C is the set of agents in the organization that play role x. Given two roles $x, y \in ROLE$ and a set of atomic actions $X \in 2^{AT}$, $\mathscr{F}_{control}(x, y) = X$ means that role x controls the atomic actions X of role y. More generally, the latter construction is used to specify a concept of right: $a \in \mathscr{F}_{control}(x, y)$ means that every agent playing role x has the *right* to require (resp. to authorize) an agent playing role y to do action a. We call the tuple $RS = \langle \mathscr{F}_{play}, \mathscr{F}_{control} \rangle$ a *role structure*.

We also introduce organizational actions of the form $req_j(a)$ and $auth_j(a)$ denoting respectively the action of requiring (or demanding) j to do the atomic action a and the action of authorizing (or allowing) j to do the atomic action a. Here we do not consider the negative counterparts of these organizational actions, that is, the action of forbidding j to do the atomic action a and the action of authorizing (or allowing) j not to do the atomic action a.

We define a set ACT of complex actions as the smallest superset of AT such that:

– if $a \in AT$ and $j \in AGT$ then $req_j(a) \in ACT$ and $auth_j(a) \in ACT$.

Since the sets AGT and AT are supposed to be finite, the set ACT is finite as well. We note $\alpha, \beta, \ldots$ the elements in ACT.

The language $\mathcal{L}_{lang}$ of the logic $\mathcal{L}$ is defined as the smallest superset of ATM such that:

– if $\varphi, \psi \in \mathcal{L}_{lang}$, $\alpha \in ACT$, $i \in AGT$, $x, y \in ROLE$ and $a \in AT$ then $\neg\varphi$, $\varphi \vee \psi$, $After_{i:\alpha}\varphi$, $Does_{i:\alpha}\varphi$, $Bel_i\varphi$, $Goal_i\varphi$, $Obg\varphi$, $Control(x, y, a)$, $Play(i, x) \in \mathcal{L}_{lang}$.

The classical boolean connectives $\wedge$, $\rightarrow$, $\leftrightarrow$, $\top$ and $\bot$ are defined from $\vee$ and $\neg$ in the usual manner.

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; $Control(x, y, a)$: role x controls role y with respect to the action a; $Play(i, x)$: agent i plays role x; $Obg\varphi$: it is obligatory that φ. During the analysis of trust presented in Section 3, formula $After_{i:\alpha}\varphi$ will be often read: agent i has the power to ensure φ by doing α. Three abbreviations are given: $Can_i(\alpha) \stackrel{\text{def}}{=} \neg After_{i:\alpha}\bot$; $Int_i(\alpha) \stackrel{\text{def}}{=} Goal_i Does_{i:\alpha}\top$; $Perm\varphi \stackrel{\text{def}}{=} \neg Obg\neg\varphi$. $Can_i(\alpha)$ 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, $Perm\varphi$ stands for: φ is permitted.

Models of the logic $\mathcal{L}$ are tuples $M = \langle W, RS, \mathscr{A}, \mathscr{D}, \mathscr{B}, \mathscr{G}, \mathscr{O}, \mathscr{V} \rangle$ defined as follows.

– W is a non empty set of possible worlds or states.
– RS is a role structure.

- $\mathscr{A} : AGT \times ACT \longrightarrow W \times W$ maps every agent i and action α to a relation $\mathscr{A}_{i:\alpha}$ between possible worlds in W. Given a world $w \in W$, if $(w, w') \in \mathscr{A}_{i:\alpha}$ then w' is a world which can be reached from w through the occurrence of agent i's action α.
- $\mathscr{D} : AGT \times ACT \longrightarrow W \times W$ maps every agent i and action α to a relation $\mathscr{D}_{i:\alpha}$ between possible worlds in W. Given a world $w \in W$, if $(w, w') \in \mathscr{D}_{i:\alpha}$ then w' is the *next* world of w which will be reached from w through the occurrence of agent i's action α.
- $\mathscr{B} : AGT \longrightarrow W \times W$ maps every agent i to a serial, transitive and euclidean relation $\mathscr{B}_i$ between possible worlds in W. Given a world $w \in W$, if $(w, w') \in \mathscr{B}_i$ then w' is a world which is compatible with agent i's beliefs at w.
- $\mathscr{G} : AGT \longrightarrow W \times W$ maps every agent i to a serial relation $\mathscr{G}_i$ between possible worlds in W. Given a world $w \in W$, if $(w, w') \in \mathscr{G}_i$ then w' is a world which is compatible with agent i's goals at w.
- $\mathscr{O}$ is a serial relation between possible worlds in W. Given a world $w \in W$, if $(w, w') \in \mathscr{O}$ then w' is a world which is ideal at world w.
- $\mathscr{V} : W \longrightarrow 2^{ATM}$ is a truth assignment which associates each world w with the set $\mathscr{V}(w)$ of atomic propositions true in w.

We distinguish the two types of relations R and D since we want to express both: the fact that at a given world w an agent performs an action α which will result in a next state w, the fact that if at w the agent did something different he would have produced a different outcome.

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 $After_{i:\alpha}\varphi$, $Does_{i:\alpha}\varphi$, $Bel_i\varphi$, $Goal_i\varphi$, $Obg\varphi$, $Control(x, y, a)$ and $Play(i, x)$.

- $M, w \models After_{i:\alpha}\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in \mathscr{A}_{i:\alpha}$
- $M, w \models Does_{i:\alpha}\varphi$ iff $\exists w'$ such that $(w, w') \in \mathscr{D}_{i:\alpha}$ and $M, w' \models \varphi$
- $M, w \models Bel_i\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in \mathscr{B}_i$
- $M, w \models Goal_i\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in \mathscr{G}_i$
- $M, w \models Obg\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in \mathscr{O}$
- $M, w \models Control(x, y, a)$ iff $a \in \mathscr{F}_{control}(x, y)$
- $M, w \models Play(i, x)$ iff $i \in \mathscr{F}_{play}(x)$

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

The axiomatizations of the logic $\mathcal{L}$ include all tautologies of propositional calculus and the standard rule of inference *modus ponens*.[1] Operators for actions of type $After_{i:\alpha}$ and $Does_{i:\alpha}$ are normal modal operators satisfying the axioms and rules of inference of system K.[2] Operators of type Bel_i and $Goal_i$ are just standard normal modal operators.

[1] If $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ then $\vdash \psi$.

[2] This includes necessitation rule and Axiom K: $\frac{\vdash \varphi}{\vdash After_{i:\alpha}\varphi}$; $\frac{\vdash \varphi}{\vdash \neg Does_{i:\alpha}\neg\varphi}$; $(After_{i:\alpha}\varphi \wedge After_{i:\alpha}(\varphi \rightarrow \psi)) \rightarrow After_{i:\alpha}\psi$; $(Does_{i:\alpha}\varphi \wedge \neg Does_{i:\alpha}\neg\psi) \rightarrow Does_{i:\alpha}(\varphi \wedge \psi)$.

The former are modal operators for belief in Hintikka style [12] satisfying the axioms and rules of inference of system KD45. [3] The latter are modal operators for goal in Cohen & Levesque's style [6] satisfying the axioms and rules of inference of system KD.[4] Thus, we make assumptions about positive and negative introspection for beliefs and we suppose that an agent have no inconsistent beliefs or conflicting goals. Operators for obligations of type Obg are supposed to be KD normal modal operators as in SDL (standard deontic logic) [1].[5] Thus, we do not admit contradictory obligations.

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

S1 if $(w, w') \in \mathscr{D}_{i:\alpha}$ and $(w, w'') \in \mathscr{D}_{j:\beta}$ 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 $(w, w') \in \mathscr{D}_{i:\alpha}$

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:

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

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).

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

S3 if $(w, w') \in \mathscr{D}_{i:\alpha}$ then $(w, w') \in \mathscr{A}_{i:\alpha}$

[3] This includes rule of necessitation, Axiom K for every operator Bel_i plus the following three axioms (so-called Axioms D, 4, 5): $\neg Bel_i\bot$; $Bel_i\varphi \rightarrow Bel_i Bel_i\varphi$; $\neg Bel_i\varphi \rightarrow Bel_i\neg Bel_i\varphi$.

[4] This includes rule of necessitation, Axiom K for every operator $Goal_i$ plus the following Axiom D: $\neg Goal_i\bot$.

[5] This includes rule of necessitation, Axiom K for Obg plus the following Axiom D: $\neg Obg\bot$.

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 $\mathbf{Inc}_{Act,PAct}$.

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

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 $\neg\varphi$ 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 [19] 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 $\forall(w, w') \in \mathscr{G}_i, \exists w''$ such that $(w', w'') \in \mathscr{D}_{i:\alpha}$ and $\exists v$ such that $(w, v) \in \mathscr{A}_{i:\alpha}$ then $\exists v'$ such that $(w, v') \in \mathscr{D}_{i:\alpha}$
S5 if $\exists v'$ such that $(w, v') \in \mathscr{D}_{i:\alpha}$ then $\forall(w, w') \in \mathscr{G}_i, \exists w''$ such that $(w', w'') \in \mathscr{D}_{i:\alpha}$

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.*[4]), 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:

S6 $\exists w'$ such that $(w, w') \in \mathscr{B}_i$ and $(w, w') \in \mathscr{G}_i$

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 $Goal_i\varphi \rightarrow Bel_i Goal_i\varphi$
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 $w \in W$:

S7 if $(w, w') \in \mathscr{B}_i$ then $\forall v$, if $(w, v) \in \mathscr{G}_i$ then $(w', v) \in \mathscr{G}_i$
S8 if $(w, w') \in \mathscr{B}_i$ then $\forall v$, if $(w', v) \in \mathscr{G}_i$ then $(w, v) \in \mathscr{G}_i$

We accept the following axiom relating obligations and beliefs:

BelObg $\quad Obg\varphi \rightarrow Bel_i Obg\varphi$

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

S9 $\quad$ if $(w, w') \in \mathscr{B}_i$ then $\forall v$, if $(w', v) \in \mathscr{O}$ then $(w, v) \in \mathscr{O}$

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: $Bel_i Perm\varphi \rightarrow Perm\varphi$. This means that in our logical framework every agent has sound information of what is permitted.

We also have specific properties for the actions of requiring and authorizing. We suppose that, given two agents i and j playing respectively roles x and y in the organization, if role x controls role y with respect to the action a then: after i requires (resp. authorizes) j to do a, j has the obligation to do a (resp. has the permission to do a). Formally:

Control $\quad (Play(i, x) \wedge Play(j, y) \wedge Control(x, y, a)) \rightarrow$
$(After_{i:req_j(a)} ObgDoes_{j:a}\top \wedge After_{i:auth_j(a)} PermDoes_{j:a}\top)$

Axiom **Control** corresponds to the following two semantic constraints over $\mathcal{L}$ models. For any $i, j \in AGT$, $x, y \in ROLE$, $a \in AT$ and $w \in W$ if $i \in \mathscr{F}_{play}(x)$, $j \in \mathscr{F}_{play}(y)$ and $a \in \mathscr{F}_{control}(x, y)$ then:

S10 $\quad$ if $(w, w') \in \mathscr{A}_{i:req_j(a)} \circ \mathscr{O}$ then $\exists w''$ such that $(w', w'') \in \mathscr{D}_{j:a}$
S11 $\quad$ if $(w, w') \in \mathscr{A}_{i:auth_j(a)}$ then $\exists w''$ such that $(w', w'') \in \mathscr{O} \circ \mathscr{D}_{j:a}$

where $\circ$ is the standard composition operator between two binary relations.

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. It is a routine task to check that the axioms of the logic $\mathcal{L}$ correspond one-to-one to their semantic counterparts on the frames. It is routine, too, to check that all of our axioms are in the Sahlqvist class. This means that the axioms are all expressible as first-order conditions on frames and that they are complete with respect to the defined frames classes, cf. [3, Th. 2.42]. □

3 Trust within the Context of Organizations

Trust relationships within the context of an organization can be analyzed at three general levels of abstraction:

- an agent's trust in another agent;
- an agent's trust in a role;
- an agent's trust in another agent *qua* player of a certain role.

The former kind of trust, also called interpersonal (or inter-agent) trust, is the trust that a certain agent i places in a different agent j. This kind of trust is based on i's ascription of specific properties to j including powers, abilities and dispositions. We call these j's *individual properties*.

On the contrary, an agent i's trust in a role x, with respect to the accomplishment of a given task φ, is based on i's attribution to role x of certain standard values and properties that are relevant for the achievement of the task φ. We call these *role properties*. For example, if i says that he trusts policemen with respect to the task of monitoring dangerous situations, i's trust in policemen is based on i's attribution to policemen of certain role properties that are relevant with respect to the task of monitoring dangerous situations (e.g. being armed, having the power to arrest suspected people, etc.).

Finally, an agent i's trust in another agent j *qua* player of a role x with respect to a certain task φ, is the trust that i places in j due to the fact that j plays role x and, according to i's beliefs, role x has certain (role) properties that are relevant for the accomplishment of task φ. In this situation, i's trust in j *qua* player of role x is based on the fact that i transfers the properties of role x (that are relevant for the accomplishment of task φ) to agent j playing role x. Differently from trust in a role, agent i's trust in agent j *qua* player of role x is also based on i's attribution to agent j of certain individual properties that are not necessarily properties of the role x. For example, i's trust in j *qua* policeman with respect to the task of monitoring dangerous situations has two facets. On the one side, it is based on the fact that j plays the role of policeman and, *qua* policeman, j inherits the role properties of policemen (e.g. being armed, having the power to arrest suspected people). On the other side, it is based on i's attribution of individual properties to j (e.g. being absent-minded and lazy). The individual properties of j might conflict with the properties that j inherits from the role of policeman leading i to negatively evaluate j with respect to the task of monitoring dangerous situations.

3.1 Interpersonal Trust

As we have stressed in our previous works [9], interpersonal trust should be 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.

In our perspective, interpersonal 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. [6]

The following is the precise concept of interpersonal trust as an *evaluation* that interests us in the present work.

Definition 1. ***Agent i's trust in agent j's action***. *Agent i trusts agent j to do α with regard to the achievement of φ if and only if i has the achievement goal that φ and i believes that:*

- *j, by doing α, will ensure φ AND*
- *j has the capacity to do α AND*
- *j intends to do α.*

The formal translation of Definition 1 is:

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

where $Trust(i, j, \alpha, \varphi)$ stands for "i trusts j to do α with regard to the achievement of φ", and formula $AGoal_i\varphi$, expressing agent i's achievement goal that φ, is defined as follows:

$$AGoal_i\varphi \stackrel{\text{def}}{=} Goal_iX\varphi \wedge \neg Bel_i\varphi$$

Our concept of achievement goal is similar to the concept studied in [6]. We say that an agent i has the achievement goal that φ if and only if, i wants φ to be true in the next state and does not believe that φ is true now. According to definition 1, i's trust in j with respect to the achievement of φ through action α is based on i's attribution of three main properties to j: the power to ensure φ by doing α ($After_{j:\alpha}\varphi$), the capacity to do action α ($Can_j(\alpha)$), the intention to do α ($Int_j(\alpha)$).

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

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

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

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

This means that i has the achievement goal of having the product:

$$AGoal_i HasProduct(i).$$

Moreover, according to i's beliefs, j, by delivering him the product, will ensure that he will have the product, and j is going to deliver the product:

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

[6] 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 him for the accomplishment of a given task. For a distinction between trust as an *evaluation* and trust as a *decision*, see [9,21].

The following theorems highlight some interesting properties of the previous notion of interpersonal trust.

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

1. $\vdash Trust(i, j, \alpha, \varphi) \rightarrow Bel_i X\varphi$
2. $\vdash Trust(i, j, \alpha, \varphi) \leftrightarrow Bel_i Trust(i, j, \alpha, \varphi)$
3. $\vdash (Trust(i, j, \alpha, \varphi) \wedge Trust(i, j, \alpha, \psi)) \rightarrow Trust(i, j, \alpha, \varphi \wedge \psi)$
4. $\vdash \neg Trust(i, j, \alpha, \top)$

Proof. We prove Theorems 2.1 and 2.4 as examples. We prove Theorem 2.1 first. $Trust(i, j, \alpha, \varphi)$ implies $Bel_i(After_{j:\alpha}\varphi \wedge Does_{j:\alpha}\top)$ (by def. of $Trust(i, j, \alpha, \varphi)$). $After_{j:\alpha}\varphi \wedge Does_{j:\alpha}\top$ implies $Does_{j:\alpha}\varphi$ (by Axiom $\mathbf{Inc}_{Act,PAct}$ and standard principles of the normal operator $Does_{j:\alpha}$). $Does_{j:\alpha}\varphi$ implies $X\varphi$ (by definition of $X\varphi$). We conclude that $Bel_i(After_{j:\alpha}\varphi \wedge Does_{j:\alpha}\top)$ implies $Bel_i X\varphi$ (by Axiom K for Bel_i).

To prove Theorem 2.4, it is sufficient to prove that $Trust(i, j, \alpha, \top)$ implies $\bot$. $Trust(i, j, \alpha, \top)$ implies $\neg Bel_i\top$ (by def. of $Trust(i, j, \alpha, \top)$ and $AGoal_i\top$). The latter implies $\bot$ (by standard principles of the normal operator Bel_i). □

According to Theorem 2.1, if i trusts j to do α with regard to φ then i has a positive expectation that φ will be true in the next state. Theorem 2.2 highlights the fact that trust is under the focus of the truster's awareness: i trusts j to do α with regard to φ if and only if, i is aware of this. Finally, Theorem 2.3 shows that trust aggregates under conjunction: if i trusts j to do α with regard to φ and i trusts j to do α with regard to ψ then, i trusts j to do α with regard to $\varphi \wedge \psi$. As Theorem 2.4 shows, in our logical model there is no trust about tautologies. This is for us an intuitive property of trust.

Trust in an Agent's Inaction. It is worth noting that an exhaustive ontology of trust must distinguish the concept *trust in an agent's action* as defined above (definition 1) from the concept of *trust in an agent's inaction*. The former concept is focused on the domain of gains whereas the latter is focused on the domain of losses. That is, in the former case the truster believes that the trustee is in condition to *further* the achievement of a pleasant state of affairs, and he will *do* that; in the latter case the truster believes that the trustee is in condition to *endanger* the maintenance of a pleasant state of affairs, but he will *refrain* from doing that. The concept of trust in an agent's inaction can be defined as follows.

Definition 2. ***Agent i's trust in agent j's inaction.*** *Agent i trusts j not to do α with regard to the maintenance of φ if and only if i has the maintenance goal that φ 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.

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

where $Trust(i, j, \neg\alpha, \varphi)$ stands for "i trusts j not to do α with regard to the maintenance of φ", and formula $MGoal_i\varphi$, expressing agent i's maintenance goal that φ, is defined as follows:

$$MGoal_i\varphi \stackrel{\text{def}}{=} Goal_iX\varphi \wedge Bel_i\varphi$$

Our concept of maintenance goal is similar to Cohen & Levesque's concept [6]: an agent i has the maintenance goal that φ if and only if, i wants φ to be true in the next state and believes that φ is true now. That is, an agent i has a maintenance goal that φ if and only if, agent i already has φ and has the goal to continue to have φ in the next state. More generally, a maintenance goal is the goal of preserving a certain state of affairs.

Example 2. Agent j is the webmaster of a public access website. 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 maintenance goal of having free access to the website:

$$MGoal_i freeAccess(i).$$

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, 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 j 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, $Trust(i, j, \neg\alpha, \varphi) \wedge \neg Bel_iX\varphi$ 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 Section 3.2 we will provide an analysis of trust in a role and trust in an agent *qua* player of a role.

3.2 Trust in a Role and Trust in an Agent *qua* Role Player

It is typical of organizations that an agent playing a certain role delegates the accomplishment of a task to another agent playing a different role. For example, an agent playing the role of director of the organization might require another agent playing the role of secretary the task of organizing a business meeting. Trust in roles plays a prominent role in organizational performance: it mediates the social interaction between agents and affects delegation mechanisms within the context of the organization [16,5].

As emphasized at the beginning of Section 3, an agent i's trust in a role x, with respect to the accomplishment of a given task φ, is based on i's attribution to role x of certain standard values and properties that are relevant for the achievement of the task φ (*role properties*). We here focus on a particular role property, that is, the (role) property of having the power to accomplish the task. In particular, we define an agent i's trust in a role x with respect to certain task as i's belief that playing role x is a sufficient

condition for an agent to have the power to accomplish the task. The precise definition of trust in a role is the following one.

Definition 3. ***Agent i's trust in role*** *x. Agent i trusts role x with regard to the achievement of φ through action α if and only if i has the achievement goal that φ and believes that:*

- *every agent playing role x, by doing α, will ensure that φ.*

The formal translation of Definition 3 is:

$$Trust(i, x, \alpha, \varphi) \stackrel{\text{def}}{=} AGoal_i\varphi \wedge Bel_i(\bigwedge_{j \in \mathscr{F}_{play}(x)} After_{j:\alpha}\varphi)$$

where $\mathscr{F}_{play}(x)$ is the set of agents which play role x in the organization. The formula $Trust(i, x, \alpha, \varphi)$ is meant to stand for "agent i trusts role x with regard to the achievement of φ through action α". The following example clarifies the meaning of the concept of trust in a role.

Example 3. Suppose that agent i is the editor in chief of a scientific journal. Agent i trusts the members of his editorial board to review an article submitted to the journal with respect to his goal of having a good evaluation of the article. Formally:

$$Trust(i, boardMember, review, goodEvaluation).$$

This means that i has the achievement goal of having a good evaluation of the article and believes that every member of the board can provide a good evaluation of the article by reviewing it:

$$AGoal_i\, goodEvaluation \wedge$$
$$Bel_i(\textstyle\bigwedge_{j \in \mathscr{F}_{play}(boardMember)} After_{j:review}\, goodEvaluation).$$

One might object that the previous definition of trust in a role x is quite strong since it requires that every agent playing role x has the power to ensure φ by doing α. One might define weaker forms of trust in a role. For instance, one might suppose that agent i trusts role x with regard to the achievement of φ through action α if and only if i has the achievement goal that φ and believes that the majority of agents playing role x can ensure φ by doing α. This alternative definition of trust in a role based on the concept of majority can be formally expressed as follows.

$$Trust(i, x, \alpha, \varphi) \stackrel{\text{def}}{=}$$

$$AGoal_i\varphi \wedge Bel_i(\bigvee_{C \subseteq \mathscr{F}_{play}(x), |C| > |\mathscr{F}_{play}(x) \setminus C|} (\bigwedge_{j \in C} After_{j:\alpha}\varphi))$$

The last kind of trust that we consider is an agent's trust in another agent *qua* player of a certain role. In our perspective, i trusts j *qua* player of role x with respect to a certain task if and only if, i trusts j *because* i thinks that j plays role x. As emphasized at the beginning of Section 3, agent i's trust in agent j *qua* player of role x has two facets. On the one side, it is based on the fact that i transfers some properties of role x (that are relevant for the accomplishment of the task) to agent j playing that role. On the other side, it is based on i's attribution of certain individual properties to j.

Definition 4. ***Agent i's trust in agent j qua player of role x.*** *Agent i trusts agent j qua player of role x with regard to the achievement of φ through action α if and only if:*

- *agent i trusts role x with regard to the achievement of φ through action α (see definition 3) AND*
- *i believes that*
  - *j plays role x AND*
  - *j has the capacity to do α AND*
  - *j intends to do α.*

According to definition 4, i's trust in j *qua* player of role x with respect to the achievement of φ through action α is based on i's trust in role x and i's attribution of two individual properties to j: the capacity to do α and the intention to do α. The definition can be formally translated as follows:

$$Trust(i, j, x, \alpha, \varphi) \stackrel{\text{def}}{=} Trust(i, x, \alpha, \varphi) \wedge Bel_i(Play(j, x) \wedge Can_j(\alpha) \wedge Int_j(\alpha))$$

where $Trust(i, j, x, \alpha, \varphi)$ stands for "agent i trusts agent j *qua* player of role x with regard to the achievement of φ through action α".

As for interpersonal trust, since in our logic the conditions $Can_j(\alpha)$ and $Int_j(\alpha)$ together are equivalent to $Does_{j:\alpha}\top$, the definition of trust in an agent *qua* player of a role can be simplified as follows:

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

Before concluding this section, we consider some formal relationships between the three concepts of trust presented above. For instance:

- is it possible that agent j plays role x and agent i trust role x with respect to the achievement of φ, without i trusting j *qua* player of role x?
- is it possible that agent i trusts agent j *qua* player of role x with respect to the achievement of φ without i trusting j?

The answer to the first question is positive. Indeed, an agent i's trust in an agent j *qua* player of a role x with respect to the achievement of φ through action α is not only based on i's trust in role x but also on i's attribution of individual properties to j (i.e. j's capacity and j's intention to do action α). Thus, it might be the case that i trusts role x, under the condition that j plays role x and, i does not trust j *qua* player of role x. This is the reason why in our logic $\mathcal{L}$ the formula $\neg Trust(i, j, x, \alpha, \varphi) \wedge Trust(i, x, \alpha, \varphi) \wedge Play(j, x)$ is satisfiable. On the contrary, the answer to the second question is negative. Indeed, it is not possible that i trusts j *qua* player of role x with respect to the achievement of φ through α and, at the same time, agent i does not trust agent j with respect to the achievement of φ through α: $Trust(i, j, x, \alpha, \varphi) \rightarrow Trust(i, j, \alpha, \varphi)$ is a theorem of the logic $\mathcal{L}$. Note also that, in our logical model, interpersonal trust does not necessarily entail trust in an agent *qua* player of a certain role, that is, i might trust j with respect to φ without trusting j *qua* player of a role with respect to φ. This is the reason why the formula $Trust(i, j, \alpha, \varphi) \wedge Play(j, x) \wedge \neg Trust(i, j, x, \alpha, \varphi)$ is satisfiable in the logic $\mathcal{L}$. This is due to the fact that i's trust in j is not generalized to all agents playing the same role as j.

In the following Section 3.3, the definitions of trust in a role and trust in an agent *qua* player of a role are applied to the specific case of an authority's trust in a subordinate.

3.3 Trust of an Authority in an Subordinate

Trust of an authority in a subordinate (e.g. the trust of a leader in a follower, of an employer in an employee, of a trainer in a player, etc.) is based on the authority's belief that the subordinate will effectively try to complete a certain delegated task, that is, an authority's trust in a subordinate is based on the authority's belief that the subordinate will conform to the obligations that the authority has created by means of certain requests.

In some of our previous papers [18] we have formally characterized the concept of *obedience* as a general attitude of the subordinate concerning norm compliance. Let us reconsider it in the context of the present analysis. We say that a certain agent i is obedient if and only if, he intends to do a certain action α as a consequence of his fulfillment of the obligation to do this action. Formally:

$$Obed_i(\alpha) \stackrel{\text{def}}{=} Bel_i ObgDoes_{i:\alpha} \top \rightarrow Int_i(\alpha)$$

where $Obed_i(\alpha)$ stands for: *i is obedient to do the action α*.

The following Theorem 3 shows how the authority's belief that the subordinate is obedient intervenes to support the authority's trust in the subordinate.

Theorem 3. *Let $i, j \in AGT$, $x, y \in ROLE$ and $a \in AT$ then:*
$\vdash (Play(i,x) \wedge Play(j,y) \wedge Control(x,y,a) \wedge$
$After_{i:req_j(a)}(Trust(i,y,\alpha,\varphi) \wedge Bel_i(Obed_j(a) \wedge Can_j(a)))) \rightarrow$
$After_{i:req_j(a)} Trust(i,j,y,a,\varphi)$

Theorems 3 has the following meaning. Suppose that agents i and j play respectively roles x and y in the organization and role x controls role y with respect to the action a. In this sense, i has authority over j with respect to the action a. Then, if after i requires j to do a, i will trust role y with respect to the achievement of φ through a and i will believe j to be capable to do a and to be obedient to do a then, after i requires j to do a, i will trust j *qua* player of role y with respect to the achievement of φ through a.

4 Conclusion

We have presented in a modal logical framework a model of trust within organizations. We have defined three different forms of trust: interpersonal trust (i.e. trust in an agent), trust in a role and trust in an agent *qua* player of a role. The formal relationships between the three concepts have been investigated. In the last part of the paper we have considered the special case of an authority's trust in a subordinate (e.g. an employer's trust in a employee). Future works will be devoted to extend our analysis to a notion of *graded trust* based on a notion of *uncertain belief*. Indeed, in the present work we have only considered a notion of *binary trust* (i.e. either i trusts j or i does not trust j). Such a kind of extension will enable us to integrate the cognitive and qualitative analysis of trust presented in this paper with a quantitative analysis and, to compare our approach with existing probabilistic approaches to trust (e.g. [15]).

References

1. Åqvist, L.: Deontic logic. In: Gabbay, D.M., Geunther, F. (eds.) Handbook of Philosophical Logic. Kluwer Academic Publishers, Dordrecht (2002)
2. Arrow, K.: The Limits of Organization. Norton, New York (1974)
3. Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge University Press, Cambridge (2001)
4. Bratman, M.: Intentions, plans, and practical reason. Harvard University Press (1987)
5. Castelfranchi, C.: Grounding organizations in the minds of agents. In: Dignum, V. (ed.) Multi-agent systems: semantics and dynamics of organizational models. IGI Global (forthcoming)
6. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artificial Intelligence 42, 213–261 (1990)
7. Coleman, J.: Foundations of Social Theory. Harvard University Press, Cambridge (1990)
8. Demolombe, R.: To trust information sources: a proposal for a modal logical framework. In: Castelfranchi, C.A., Tan, Y.H. (eds.) Trust and Deception in Virtual Societies, pp. 111–124. Kluwer, Dordrecht (2001)
9. Falcone, R., Castelfranchi, C.: Social trust: A cognitive approach. In: Castelfranchi, C., Tan, Y.H. (eds.) Trust and Deception in Virtual Societies, pp. 55–90. Kluwer, Dordrecht (2001)
10. Grossi, D., Royakkers, L., Dignum, F.: Organizational structure and responsibility: an analysis in a dynamic logic of organized collective agency. Artificial Intelligence and Law 15, 223–249 (2007)
11. Harel, D., Kozen, D., Tiuryn, J.: Dynamic Logic. MIT Press, Cambridge (2000)
12. Hintikka, J.: Knowledge and Belief. Cornell University Press, New York (1962)
13. Hübner, J.F., Sichman, J.S., Boissier, O.: Developing organised multi-agent systems using the MOISE+ model: programming issues at the system and agent levels. International Journal of Agent-Oriented Software Engineering 1(3/4), 370–395 (2007)
14. Jones, A.J.I.: On the concept of trust. Decision Support Systems 33(3), 225–232 (2002)
15. Jøsang, A.: A logic for uncertain probabilities. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9(3), 279–311 (2001)
16. Kramer, R.M.: Trust and distrust in organizations: emerging perspectives, enduring questions. Annual Review of Psychology 50, 569–598 (1999)
17. Liau, C.J.: Belief, information acquisition, and trust in multi-agent systems: a modal logic formulation. Artificial Intelligence 149, 31–60 (2003)
18. Lorini, E., Demolombe, R.: Trust and norms in the context of computer security: A logical formalization. In: van der Meyden, R., van der Torre, L. (eds.) DEON 2008. LNCS, vol. 5076, pp. 50–64. Springer, Heidelberg (2008)
19. Lorini, E., Herzig, A.: A logic of intention and attempt. Synthese 163(1), 45–77
20. March, J.G., Simon, H.: Organizations. John Wiley & Sons, Chichester (1958)
21. Marsh, S.: Formalising Trust as a Computational Concept. PhD thesis, University of Stirling, Scotland (1994)
22. Santos, F., Carmo, J., Jones, A.: Action concepts for describing organised interaction. In: Sprague, R.A. (ed.) Thirtieth Annual Hawai International Conference on System Sciences, pp. 373–382. IEEE Computer Society Press, Los Alamitos (1997)
23. Wooldridge, M., Jennings, N.R., Kinny, D.: The Gaia methodology for agent-oriented analysis and design. Autonomous Agents and Multi-Agent Systems 3(3), 285–312 (2000)