From Binary Trust to Graded Trust in Information Sources: A Logical Perspective

Emiliano Lorini, Robert Demolombe

From Binary Trust to Graded Trust in Information Sources: A Logical Perspective

Emiliano Lorini and Robert Demolombe

Institut de Recherche en Informatique de Toulouse (IRIT), France
Emiliano.Lorini@irit.fr
Robert.Demolombe@irit.fr

Abstract. We present a concept of trust that integrates the truster's goal, the trustee's action that ensures the achievement of the truster's goal, and the trustee's ability and intention to do this action. This concept of trust is formalized in modal logic and is applied to the particular domain of trust in information sources. In this context trust may be derived, in particular, from the truster's beliefs about some properties of the information source: validity, completeness, sincerity, competence, vigilance and cooperativity. In the last part of the paper we move beyond binary trust (i.e. either i trusts j or i does not trust j) in order to capture a concept of graded trust.

1 Introduction

Trust in information sources plays a crucial role in many areas of interaction between agents, in particular when information sources are software agents. A typical example is in the field of stock and bond market where trust has a strong influence on a decision to buy, or to sale, a specific kind of stocks. To take such decisions agents have several types of information sources to consult in order to predict the future evolution of the stock value. These information sources may be banks, companies, consultants, etc. and the agents may believe that some of these sources have a good competence but are not necessarily sincere, others are reluctant to inform about bad news, others are competent but are not necessarily informed at the right moment, etc.

We think that reasoning about so complex situations requires a clear definition of trust and of its main dimensions, and safe inference rules that can be applied by the agents. The aim of this work is to present a formal model of trust which meets the previous desiderata. We will first present and formalize a general concept of trust as the truster's *evaluation* of specific properties of the trustee (powers, abilities, dispositions) which are together sufficient to ensure that a goal of the truster will be achieved. Then, we will apply this general concept of trust to the specific case of trust in information sources. It is worth noting that it is out of the scope of our work to propose a model of trust based on statistics about past interactions with a given target and reputational information.

The paper is organized as follows. We start with a presentation of a logical framework which is used for formalizing the relevant concepts of the present analysis of trust in information sources (Section 2). In Section 3 a general definition of trust is presented

and its main properties are discussed. In the second part of the paper we start with a formal characterization of the properties of information sources: validity, completeness, sincerity, competence, vigilance and cooperativity (Section 4). We show that these properties are epistemic supports for trust in information sources (Section 5). In the last part of the paper (Section 6) we show how the logical framework presented in Section 2 can be appropriately extended in order to move beyond an analysis of binary trust (i.e. either agent i trusts agent j or agent i does not trust agent j) and to capture a concept of graded trust (i.e. agent i trusts agent j with a certain strength k).

2 Setting Up the Formalism

We present in this section the multimodal logic $\mathcal{L}$ that we use in the paper to formalize the relevant concepts of our model of trust. $\mathcal{L}$ combines the expressiveness of dynamic logic [17] with the expressiveness of a so-called BDI (belief, desire, intention) logic of agents' mental attitudes (see [7] for instance).

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 nonempty finite set of atomic formulas $\Pi = \{p, q, \ldots\}$.

LIT is the set of literals which includes all atomic formulas and their negations, that is:

- $LIT = \{p, \neg p | p \in \Pi\}$.

We note $P, Q, \ldots$ the elements in LIT. We also introduce specific actions of the form $inf_j(P)$ denoting the action of informing agent j that P is true. We call them informative actions. The set $INFO$ of informative actions is defined as follows:

- $INFO = \{inf_j(P) | j \in AGT,\ P \in LIT\}$.

Since the set Π is finite, the set $INFO$ is finite as well. The set ACT of complex actions is given by the union of the set of atomic actions and the set of informative actions, that is:

- $ACT = AT \cup INFO$.

We note $\alpha, \beta, \ldots$ the elements in ACT. The language of $\mathcal{L}$ is the set of formulas defined by the following BNF:

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

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

The operators of our logic have the following intuitive meaning. $Bel_i\varphi$: the agent i believes that φ; $After_{i:\alpha}\varphi$: immediately after agent i does α, it is the case that φ ($After_{i:\alpha}\bot$ is read: agent i cannot do action α); $Does_{i:\alpha}\varphi$: agent i is going to do α and

φ will be true afterward ($Does_{i:\alpha}\top$ is read: agent i is going to do α); $Goal_i\varphi$: the agent i wants that φ holds.

The following abbreviations are given:

$$Can_i(\alpha) \stackrel{\text{def}}{=} \neg After_{i:\alpha}\bot$$

$$Int_i(\alpha) \stackrel{\text{def}}{=} Goal_i Does_{i:\alpha}\top$$

$$Inf_{i,j}(P) \stackrel{\text{def}}{=} Does_{i:inf_j(P)}\top$$

$Can_i(\alpha)$ stands for: agent i can do action α (i.e. i has the capacity to do α). $Int_i(\alpha)$ stands for: agent i intends to do α. Finally $Inf_{i,j}(P)$ stands for: i informs j that P is true.

Models of the logic $\mathcal{L}$ ($\mathcal{L}$ models) are tuples $M = \langle W, R, D, B, G, V\rangle$ defined as follows.

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

We distinguish here two types of relations for specifying the dynamic dimension of models (relations of type R and type 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 $After_{i:\alpha}\varphi$, $Does_{i:\alpha}\varphi$, $Bel_i\varphi$ and $Goal_i\varphi$.

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

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

2.2 Axiomatization

The axiomatization of the logic $\mathcal{L}$ includes all tautologies of propositional calculus and the rule of inference *modus ponens* (**MP**).

MP From $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ infer $\vdash \psi$

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 [6]. Operators of type Bel_i and $Goal_i$ are just standard normal modal operators. The former are modal operators for belief in Hintikka style [19] satisfying the axioms and rules of inference of system KD45, whereas the latter are modal operators for goal in Cohen & Levesque's style [7] satisfying the axioms and rules of inference of system KD. That is, the following axioms and rules of inference for every operator Bel_i, $Goal_i$, $After_{i:\alpha}$ and $Does_{i:\alpha}$ are given.

$\mathbf{K}_{Bel}$ $(Bel_i\varphi \wedge Bel_i(\varphi \rightarrow \psi)) \rightarrow Bel_i\psi$
$\mathbf{K}_{Goal}$ $(Goal_i\varphi \wedge Goal_i(\varphi \rightarrow \psi)) \rightarrow Goal_i\psi$
$\mathbf{K}_{PAct}$ $(After_{i:\alpha}\varphi \wedge After_{i:\alpha}(\varphi \rightarrow \psi)) \rightarrow After_{i:\alpha}\psi$
$\mathbf{K}_{Act}$ $(Does_{i:\alpha}\varphi \wedge \neg Does_{i:\alpha}\neg\psi) \rightarrow Does_{i:\alpha}(\varphi \wedge \psi)$
$\mathbf{D}_{Bel}$ $\neg(Bel_i\varphi \wedge Bel_i\neg\varphi)$
$\mathbf{D}_{Goal}$ $\neg(Goal_i\varphi \wedge Goal_i\neg\varphi)$
$\mathbf{4}_{Bel}$ $Bel_i\varphi \rightarrow Bel_iBel_i\varphi$
$\mathbf{5}_{Bel}$ $\neg Bel_i\varphi \rightarrow Bel_i\neg Bel_i\varphi$
$\mathbf{Nec}_{Bel}$ From $\vdash \varphi$ infer $\vdash Bel_i\varphi$
$\mathbf{Nec}_{Goal}$ From $\vdash \varphi$ infer $\vdash Goal_i\varphi$
$\mathbf{Nec}_{PAct}$ From $\vdash \varphi$ infer $\vdash After_{i:\alpha}\varphi$
$\mathbf{Nec}_{Act}$ From $\vdash \varphi$ infer $\vdash \neg Does_{i:\alpha}\neg\varphi$

Axioms $\mathbf{K}_{Bel}$ and $\mathbf{K}_{Goal}$ with rules of inference $\mathbf{Nec}_{Bel}$ and $\mathbf{Nec}_{Goal}$ are the principles of a minimal normal modal logic for every modal operator Bel_i and every modal operator $Goal_i$. Axioms $\mathbf{K}_{PAct}$ and $\mathbf{K}_{Act}$ with rules of inference $\mathbf{Nec}_{PAct}$ and $\mathbf{Nec}_{Act}$ are the principles of a minimal normal modal logic for every modal operator $After_{i:\alpha}$ and every modal operator $Does_{i:\alpha}$. Axioms $\mathbf{D}_{Bel}$, $\mathbf{4}_{Bel}$, $\mathbf{5}_{Bel}$ correspond (in the sense of correspondence theory, see for instance [35,2]) to the seriality, transitivity and euclideanity of every relation B_i, whereas Axiom $\mathbf{D}_{Goal}$ corresponds to the seriality of every relation G_i. Thus, we suppose positive and negative introspection over beliefs (Axioms $\mathbf{4}_{Bel}$ and $\mathbf{5}_{Bel}$) and we assume that an agent cannot have inconsistent beliefs and conflicting goals (Axioms $\mathbf{D}_{Bel}$ and $\mathbf{D}_{Goal}$).

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

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

Constraint $S1$ says that if w' is the *next* world of w which is reachable from w through the occurrence of agent i's action α and w'' is also the *next* world of w which is reachable from w through the occurrence of agent j's action β, then w' and w'' denote the

same world. Indeed, we suppose that every world can only have one *next* world. The semantic constraint $S1$ corresponds to the following axiom.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 [29,30] 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 $w \in W$:

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

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

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

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

S6 $\exists w'$ such that $(w,w') \in B_i$ and $(w,w') \in G_i$

The constraint $S6$ 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 $(w,w') \in B_i$ and $(w',v) \in G_i$ then $(w,v) \in G_i$
S8 if $(w,w') \in B_i$ and $(w,v) \in G_i$ then $(w',v) \in G_i$

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

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

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

S9 if $(w,w') \in R_{j:\alpha} \circ B_i$ and $\exists v$ such that $(w,v) \in B_i \circ R_{j:\alpha}$ then $(w,w') \in B_i \circ R_{j:\alpha}$

where $\circ$ is the standard composition operator between two binary relations. In accepting the Axiom **NF**, we suppose that events are always uninformative, that is, i should not forget anything about the particular effects of j's action α that starts at a world w. What an agent i believes at a world w', only depends on what i believed at the previous world w and on the action which has occurred and which was responsible for the transition

[2] See also [13,15,18,33] for a discussion of this property.

from w to w'. Besides, Axiom **NF** relies on an additional assumption of complete and correct information. It is supposed that j's action α occurs if and only if every agent is informed of this fact. Hence all action occurrences are supposed to be public.

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

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

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

S10 if $\exists w'$ such that $(w, w') \in D_{j:inf_i(P)}$ then
$\forall (w, v) \in B_i, \exists w''$ such that $(v, w'') \in D_{j:inf_i(P)}$
S11 if $\exists w', w''$ such that $(w, w') \in B_i$ and $(w', w'') \in D_{j:inf_i(P)}$ then $\exists v$ such that $(w, v) \in D_{j:inf_i(P)}$

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

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

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

3 A General Definition of Trust

In this work trust is conceived as a complex configuration of mental states in which there is both a motivational component and a doxastic component. More precisely, we assume that i's trust in agent j necessarily involves a goal of the truster: if 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 i trusts agent j then necessarily i trusts j because i has some goal and believes that j has the right properties to ensure that such a goal will be achieved.

The concept of trust formalized in this work is similar to the concept of trust defined by Castelfranchi & Falcone [5,14]. We agree with them that trust should not be seen as an unitary and simplistic notion as other models implicitly suppose. For instance, there are computational models of trust in which trust is conceived as an expectation of the truster about a successful performance of the trustee sustained by the repeated direct interactions with the trustee (under the assumption that iterated experiences of success strengthen the truster's confidence) [23,37]. More sophisticated models of social trust

have been developed in which reputational information is added to information obtained via direct interaction (e.g. [20,32]). All these models are in our view over-simplified since they do not consider the beliefs supporting the truster's expectation which enter into play in the truster's evaluation of the trustee.

On this point we agree with Castelfranchi & Falcone on the fact that: trust is based on the truster's ascription of specific properties to the trustee (e.g. abilities, competencies, dispositions, etc.) and to the environment in which the trustee is going to act, which are together sufficient to ensure that the truster will achieve one of his goals. In this perspective, trust is nothing more than the truster's *evaluation* of certain relevant properties of the trustee.[3]

Here we just focus on a particular form of trust that can be called *trust in the trustee's action*. According to the proposed definition, agent i trusts agent j to do a certain action α if and only if i has a certain goal and thinks that j will perform action α in such a way that his goal will be achieved.[4]

The concept of trust we are interested in here is the following.

Definition 1. ***TRUST IN THE TRUSTEE'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:*

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

The three conditions 1, 2 and 3 can reformulated in formal terms as follows.

- Condition $C1$: $After_{j:\alpha}\varphi$
- Condition $C2$: $Can_j(\alpha)$
- Condition $C3$: $Int_j(\alpha)$

Condition $C1$ concerns the trustee's power to satisfy the truster's goal that φ by means of the performance of action α. Conditions $C2$ and $C3$ are about the trustee's properties which are necessary and sufficient for him to perform action 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:

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

[3] In this paper we do not consider a related notion of *decision to trust*, that is, the truster's decision to bet and wager on the trustee and to rely on him for the accomplishment of a given task. For a distinction between trust as an *evaluation* and trust as a *decision*, see [14,31].

[4] In a complementary work [28] we have provided a richer typology of trust by distinguishing *trust in the trustee's action* from *trust in the trustee's inaction*. This opposition is symmetrical to the opposition between *doing* and *refraining* (or *forbearing*) which has been studied in the philosophy of action [1,36].

Our concept of achievement goal is similar to the concept studied in [7]. 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.

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 $\mathbf{Inc}_{Act,PAct}$, **IntAct1** and **IntAct2**), so the definition of trust in the trustee's action can be simplified as follows:

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

This formalization of definition 1 better expresses a fundamental aspect of the concept of trust, namely the fact that the truster has an achievement goal that φ and believes that the trustee will ensure φ by doing action α.

Example 1. Suppose that Bill trusts Mary to shoot Bob with regard to his goal that Bob will die in the next state:

$$Trust(Bill, Mary, shoot, \neg BobAlive).$$

This means that Bill has the achievement goal that Bob will die in the next state:

$$AGoal_{Bill}\neg BobAlive.$$

Moreover, according to Bill's beliefs, Mary, by shooting Bob, will ensure that Bob is dead in the next state, and Mary is going to shoot Bob:

$$Bel_{Bill}(After_{Mary:shoot}\neg BobAlive \wedge Does_{Mary:shoot}\top).$$

The following theorems highlight some interesting properties of the previous notion of 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 2.2 as examples. We start with Theorem 2.1. $Trust(i,j,\alpha,\varphi)$ implies $Bel_i(After_{j:\alpha}\varphi \wedge Does_{j:\alpha}\top)$ (by definition 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 stardard 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 $\mathbf{K}_{Bel}$).

Let us consider Theorem 2.2. $Trust(i,j,\alpha,\varphi)$ is equivalent to $Goal_i X\varphi \wedge \neg Bel_i\varphi \wedge Bel_i(After_{j:\alpha}\varphi \wedge Does_{j:\alpha}\top)$, by definition of $Trust(i,j,\alpha,\varphi)$. The latter implies $Bel_i(Goal_i X\varphi \wedge \neg Bel_i\varphi \wedge Bel_i(After_{j:\alpha}\varphi \wedge Does_{j:\alpha}\top))$ (by Axioms $\mathbf{4}_{Bel}$, $\mathbf{5}_{Bel}$ and **PIntrGoal**) which in turn implies $Bel_i Trust(i,j,\alpha,\varphi)$ (by definition of $Trust(i,j,\alpha,\varphi)$). The other direction of Theorem 2.2 is provable by the same principles. □

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.

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

4 Basic Properties of an Information Source

We suppose that the properties of an information source can be defined in terms of the relationships between three facts:

- an information source j informs an agent i that a certain fact P is true;
- an information source j believes that P is true;
- the fact P is true.

The properties of information sources can be all expressed in a conditional form. The systematic analysis of these relationships between the previous three facts leads to six different properties of information sources.

Definition 2. ***INFORMATION SOURCE VALIDITY.*** *Agent j is a valid information source about P with regard to i if and only if, after j does the action of informing i about P, it is the case that P.*

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

Note that $After_{j:inf_i(P)}P$ can be read in an explicit conditional form: if j has the capacity to do the action of informing i about P then, P is true after every occurrence of this action. Indeed, $After_{j:inf_i(P)}P$ is logically equivalent to $Can_j(inf_i(P)) \rightarrow After_{j:inf_i(P)}P$.

Definition 3. ***INFORMATION SOURCE COMPLETENESS.*** *Agent j is a complete information source about P with regard to i if and only if, if P is true then j does the action of informing i about P.*

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

Definition 4. ***INFORMATION SOURCE SINCERITY.*** *Agent j is a sincere information source about P with regard to i if and only if, after j does the action of informing i about P, it is the case that j believes P.*

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

Note that $After_{j:inf_i(P)}Bel_jP$ too can be read in an explicit conditional form: if j has the capacity to do the action of informing i about P then, j believes P after every occurrence of this action. Indeed, $After_{j:inf_i(P)}Bel_jP$ is logically equivalent to $Can_j(inf_i(P)) \rightarrow After_{j:inf_i(P)}Bel_jP$.

Definition 5. ***INFORMATION SOURCE COMPETENCE.*** *Agent j is a competent information source about P if and only if, if j believes P then P is true.*

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

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

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

Definition 7. ***INFORMATION SOURCE COOPERATIVITY.*** *Agent j is a cooperative information source about P with regard to i if and only if, if j believes that P then j informs i about P.* [5]

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

The previous properties of information sources are not independent. For instance, as the following Theorems 3.1 and 3.2 show, validity can be derived from sincerity and competence, and completeness can be derived from vigilance and cooperativity.

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

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

Proof. We give the proof of Theorem 3.1 as an example.

$Sinc(j, i, P) \wedge After_{j:inf_i(P)} Compet(j, P)$ is equivalent to $After_{j:inf_i(P)} Bel_j P \wedge After_{j:inf_i(P)}(Bel_j P \rightarrow P)$ (by definitions of $Compet(j, P)$ and $Sinc(j, i, P)$). The latter imply $After_{j:inf_i(P)} P$ (by Axiom $\mathbf{K}_{PAct}$) which is equivalent to $Valid(j, i, P)$. □

Note that in Theorem 3.1, the derivation of $Valid(j, i, P)$ requires j's competence at the instant where the action $inf_i(P)$ has been performed by j. This is the reason why we have $After_{j:inf_i(P)} Compet(j, P)$ in the antecedent.

Example 2. Consider an example in the field of stocks and bonds market. The agent BUG is the Bank of Union of Groenland. Sue Naive (SN) and Very Wise (VW) are two BUG's customers. BUG plays the role of an information source for the customers, for instance for the facts p: "it is recommended to buy MicroHard stocks", and q: "Microhard stocks are dropping". SN believes that BUG is sincere with regard to her about p and BUG is competent about p, because SN believes that BUG wants to help its customers and BUG has a long experience in the domain. SN also believes that BUG is cooperative with regard to her about q because q is a relevant information for customers in order to make decisions. VW too believes that BUG is competent about p. But VW does not believe that BUG is sincere with regard to him about p. Indeed, VW believes that BUG wants that VW buys Microhard stocks, even if this is not profitable for VW. This example is formally represented by the following formula: $Bel_{SN} Sinc(BUG, SN, p) \wedge Bel_{SN} Compet(BUG, p) \wedge Bel_{SN} Coop(BUG, SN, q) \wedge$

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

[5] This definition of cooperativity does not exclude that i does not want to be informed about P, like in spamming.

5 Trust in Information Sources

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

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

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

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

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

Proof. $Bel_i(Compl(j,i,p) \wedge Compl(j,i,\neg p))$ implies $Bel_i((p \rightarrow Inf_{j,i}(p)) \wedge (\neg p \rightarrow Inf_{j,i}(\neg p)))$ (by definitions of $Compl(j,i,p)$ and $Compl(j,i,\neg p)$). The latter implies $Bel_i((p \rightarrow Inf_{j,i}(p)) \wedge (\neg p \rightarrow Inf_{j,i}(\neg p)) \wedge (p \vee \neg p))$ which in turn implies $Bel_i(Inf_{j,i}(p) \vee Inf_{j,i}(\neg p))$ (by standard principles of propositional calculus). Since $Inf_{j,i}(P)$ is equivalent to $Bel_iInf_{j,i}(P)$ (by Axioms **PIntrInf**, **NIntrInf** and $\mathbf{D}_{Bel}$), $Bel_i(Inf_{j,i}(p) \vee Inf_{j,i}(\neg p))$ is equivalent to $Bel_i(Bel_iInf_{j,i}(p) \vee Bel_iInf_{j,i}(\neg p))$. Let us now prove that $Bel_i(Bel_iInf_{j,i}(p) \vee Bel_iInf_{j,i}(\neg p))$ implies $Bel_iInf_{j,i}(p) \vee Bel_iInf_{j,i}(\neg p)$. $Bel_i(Bel_iInf_{j,i}(p) \vee Bel_iInf_{j,i}(\neg p))$ implies $\neg Bel_i(\neg Bel_iInf_{j,i}(p) \wedge \neg Bel_iInf_{j,i}(\neg p))$ (by Axiom $\mathbf{D}_{Bel}$). By standard principles of the normal operator Bel_i, the latter implies $\neg Bel_i\neg Bel_iInf_{j,i}(p) \vee \neg Bel_i\neg Bel_iInf_{j,i}(\neg p)$ which in turn implies $Bel_iInf_{j,i}(p) \vee Bel_iInf_{j,i}(\neg p)$ (by Axiom $\mathbf{5}_{Bel}$).

We can conclude that $Bel_i(Compl(j,i,p) \wedge Compl(j,i,\neg p))$ implies $Bel_iInf_{j,i}(p) \vee Bel_iInf_{j,i}(\neg p)$. Thus, $Bel_i(Valid(j,i,p) \wedge Valid(j,i,\neg p)) \wedge Bel_i(Compl(j,i,p) \wedge Compl(j,i,\neg p)) \wedge AGoal_iKW_ip$ implies $Bel_i(Valid(j,i,p) \wedge$

$Valid(j,i,\neg p)) \wedge \quad (Bel_i Inf_{j,i}(p) \vee Bel_i Inf_{j,i}(\neg p)) \wedge \quad AGoal_i KW_i p$ which in turn implies $(Bel_i Valid(j,i,p) \wedge Bel_i Inf_{j,i}(p) \wedge AGoal_i KW_i p) \vee (Bel_i Valid(j,i,\neg p) \wedge Bel_i Inf_{j,i}(\neg p) \wedge AGoal_i KW_i p)$. Now, let us prove that the latter implies $Trust(i,j,inf_i(p), KW_i p) \vee Trust(i,j,inf_i(\neg p), KW_i p)$.

We distinguish two cases. First, we prove that $Bel_i Valid(j,i,p) \wedge Bel_i Inf_{j,i}(p) \wedge AGoal_i KW_i p$ implies $Trust(i,j,inf_i(p), KW_i p)$. $Bel_i Valid(j,i,p) \wedge Bel_i Inf_{j,i}(p) \wedge AGoal_i KW_i p$ implies $Bel_i After_{j:inf_i(P)} P \wedge Bel_i Inf_{j,i}(p) \wedge AGoal_i KW_i p$ (by definition of $Valid(j,i,p)$) which in turn implies $Bel_i After_{j:inf_i(P)} P \wedge Bel_i Inf_{j,i}(p) \wedge \neg Bel_i \neg Can_j(inf_i(P)) \wedge AGoal_i KW_i p$ (by Axioms $\mathbf{Inc}_{Act,PAct}$, $\mathbf{K}_{Bel}$, $\mathbf{D}_{Bel}$, and definitions of $Inf_{j,i}(p)$ and $Can_j(inf_i(P))$). From the latter and by Axioms $\mathbf{4}_{Bel}$ and $\mathbf{5}_{Bel}$, we can infer $Bel_i After_{j:inf_i(P)} P \wedge Bel_i Inf_{j,i}(p) \wedge Bel_i \neg Bel_i \neg Can_j(inf_i(P)) \wedge Bel_i Bel_i After_{j:inf_i(P)} P \wedge AGoal_i KW_i p$. The latter implies $Bel_i After_{j:inf_i(P)} P \wedge Bel_i Inf_{j,i}(p) \wedge Bel_i After_{j:inf_i(P)} Bel_i P \wedge AGoal_i KW_i p$ (by standard principles of the normal operator Bel_i, Axioms $\mathbf{K}_{Bel}$ and **NF**) which in turn implies $Bel_i Inf_{j,i}(p) \wedge Bel_i After_{j:inf_i(P)} KW_i p \wedge AGoal_i KW_i p$ (by Axiom $\mathbf{K}_{Bel}$, definition of $KW_i p$, standard principles of the normal operators Bel_i and $After_{j:inf_i(P)}$). The latter is equivalent to $Trust(i,j,inf_i(p), KW_i p)$. In a similar way we can prove that $Bel_i Valid(j,i,\neg p) \wedge Bel_i Inf_{j,i}(\neg p) \wedge AGoal_i KW_i p$ implies $Trust(i,j,inf_i(\neg p), KW_i p)$.

This is sufficient to prove Theorem 4. □

From Theorems 3.1 and 3.2, similar theorems can be proved by substituting $Valid(j,i,p)$ with $(Sinc(j,i,p) \wedge After_{j:inf_i(p)} Compet(j,p))$, $Valid(j,i,\neg p)$ with $(Sinc(j,i,\neg p) \wedge After_{j:inf_i(\neg p)} Compet(j,\neg p))$, $Compl(j,i,p)$ with $Vigil(j,p) \wedge Coop(j,i,p)$ and, $Compl(j,i,\neg p)$ with $Vigil(j,\neg p) \wedge Coop(j,i,\neg p)$ in the antecedent of Theorem 4.

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

Theorem 5. *Let $i,j \in AGT$ and $inf_i(p), inf_i(\neg p) \in INFO$, then:*

$\vdash (Trust(i,j,inf_i(p), KW_i p) \vee Trust(i,j,inf_i(\neg p), KW_i p)) \rightarrow Bel_i X KW_i p$

Example 3. Let us consider again the example of stocks and bonds market. SN has the epistemic goal of knowing whether q ("Microhard stocks are dropping") is true:

$$AGoal_{SN}\ KW_{SN} q.$$

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

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

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

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

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

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

6 The Graded Aspect of Trust

In this section we show how the logic presented in Section 2 can be appropriately modified by substituting the doxastic operators of type Bel_i with operators of graded beliefs which enable to express that an agent believes some fact φ with strength k. We will exploit these operators in order to enrich Definition 1 of trust in the trustee's action, moving from *binary trust* (i.e. either i trusts j or i does not trust j) to *graded trust* (i.e. agent i trusts agent j with a certain strength k).

6.1 A Logic of Graded Beliefs

We introduce a a non-empty set of n integers $I = \{1, \ldots, n\}$. For every integer $x \in I$ and agent $i \in AGT$ we suppose a corresponding operator $Bel_i^{\geq x}$. Operators of type $Bel_i^{\geq x}$ substitute the doxastic operators Bel_i in Hintikka's style [19] introduced in Section 2. A formula $Bel_i^{\geq x}\varphi$ is meant to stand for "agent i believes φ at least with strength x". We sketch the semantics of these operators and the corresponding axiomatization.

Every operator $Bel_i^{\geq x}$ is interpreted according to a binary relation P_i^x between possible worlds in W. These binary relations are used to specify the degree of *exceptionality* of a certain world according to a given agent.[6] The binary relations P_i^x substitute the binary relations B_i as defined in Section 2. Thus, the definition of $\mathcal{L}$ models has to be modified accordingly. Given an arbitrary world $w \in W$, if $(w, w') \in P_i^x$ then w' is a world that at world w is *exceptional* for agent i at most with degree x. For every $i \in AGT$ and $x \in I$, we note $P_i^x(w) = \{w' : (w, w') \in P_i^x\}$ the set of worlds that at world w are exceptional for agent i at most with degree x.

The truth definition of formula $Bel_i^{\geq x}\varphi$ is given as follows.

- $M, w \models Bel_i^{\geq x}\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in P_i^x$

As far as the semantic constraints are concerned, we suppose that, for any $x, y \in I$ such that $y < x$, the set of worlds that at world w are for agent i exceptional at most with degree y is a subset of the set of worlds that are for agent i exceptional at most with degree x. Formally, for any $w \in W$, $i \in AGT$ and $x, y \in I$ such that $y < x$:

S12 if $(w, w') \in P_i^y$ then $(w, w') \in P_i^x$.

[6] Note that the *exceptionality* degree of a certain world should be considered as the opposite of its *possibility* degree. That is, the *exceptionality* degree of a certain world decreases when its *possibility* degree increases.

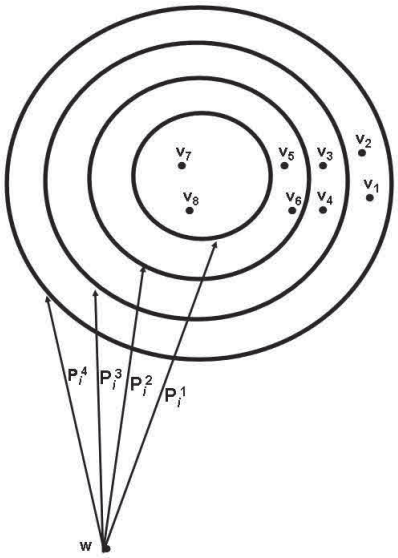

Fig. 1.

In this sense, worlds in a model are ordered according to their exceptionality degrees and form a "system of spheres" [26]. In particular, an *ordinal conditional function* (OCF) $\kappa_i^w : W \longmapsto I$ in Spohn's sense [34] can be associated with every agent i and every world w in a model M, where $\kappa_i^w(w')$ corresponds to the degree of exceptionality of world w' for agent i at world w.[7] Every function κ_i^w can be defined from the relation P_i as follows:

- $\kappa_i^w(w') = 1$ if and only if $w' \in P_i^1(w)$;
- for every $x \in I$ such that $x > 1$, $\kappa_i^w(w') = x$ if and only if $w' \in P_i^x(w)$ and $w' \notin P_i^{x-1}(w)$.

By way of example, suppose that there are four different exceptionality degrees as in Fig. 1, that is, $I = \{1, 2, 3, 4\}$. The eight worlds v_1-v_8 are ordered as follows:

- $P_i^4(w) = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8\}$;
- $P_i^3(w) = \{v_3, v_4, v_5, v_6, v_7, v_8\}$;
- $P_i^2(w) = \{v_5, v_6, v_7, v_8\}$;
- $P_i^1(w) = \{v_7, v_8\}$.

This means that worlds v_1-v_8 are the worlds that are exceptional for agent i at most with maximal degree 4; worlds v_3-v_8 are the worlds that are exceptional for agent i at most with degree 3; worlds v_5-v_8 are the worlds that are exceptional for agent i at most with degree 2; worlds v_7 and v_8 are the two worlds that are exceptional for agent i at most with minimal degree 1.

[7] On this issue, see also [25].

The previous four items contain even more information. Worlds v_1 and v_2 belong to $P_i^4(w)$ and do not belong to $P_i^3(w)$, hence they are worlds that are for agent i exceptional with degree 4. That is, $\kappa_i^w(v_1) = \kappa_i^w(v_2) = 4$. Worlds v_3 and v_4 belong to $P_i^3(w)$ and do not belong to $P_i^2(w)$, hence they are worlds that are for agent i exceptional with degree 3. That is, $\kappa_i^w(v_3) = \kappa_i^w(v_4) = 3$. Worlds v_5 and v_6 belong to $P_i^2(w)$ and do not belong to $P_i^1(w)$, hence they are worlds that are for agent i exceptional with degree 2. That is, $\kappa_i^w(v_5) = \kappa_i^w(v_6) = 2$. Finally, v_7 and v_8 belong to $P_i^1(w)$, hence they are worlds that are for agent i exceptional with degree 1. That is, $\kappa_i^w(v_7) = \kappa_i^w(v_8) = 1$.

The following is the logical axiom which corresponds to the previous semantic constraint $S12$. For any, $x, y \in I$ such that $y < x$:

Inc$_{x,y}$ $\quad Bel_i^{\geq x}\varphi \rightarrow Bel_i^{\geq y}\varphi$

According to Axiom **Inc**$_{x,y}$, if agent i believes φ at least with strength x and $y < x$ then, agent i believes φ at least with strength y.

Additional reasonable principles for operators of type $Bel_i^{\geq x}$ are the following. Suppose that x is the bigger number in I, that is, $x, y \in I$ and $\forall z \in I$, $z \geq x$. Then:

D$_{max}$	$\neg(Bel_i^{\geq x}\varphi \wedge Bel_i^{\geq x}\neg\varphi)$
PIntrPoss	$Bel_i^{\geq z}\varphi \rightarrow Bel_i^{\geq x}Bel_i^{\geq z}\varphi$
NIntrPoss	$\neg Bel_i^{\geq z}\varphi \rightarrow Bel_i^{\geq x}\neg Bel_i^{\geq z}\varphi$

According to Axiom **D**$_{max}$, it is never the case that an agent believes both φ and $\neg\varphi$ with maximal strength x. According to Axiom **PIntrPoss** and Axiom **NIntrPoss**, if an agent believes (resp. does not believe) φ at least with strength z then, he believes with maximal strength x that he believes (resp. does not believe) φ at least with strength z.

The three axioms correspond to the following three first-order properties of Kripke models. Suppose that $x, y \in I$ and $\forall z \in I$, $z \geq x$. Then, for any $w \in W$, $i \in AGT$ and $z \in I$:

S13	$\exists w' \in W$ such that $(w, w') \in P_i^x$
S14	if $(w, w') \in P_i^x$ and $(w', v) \in P_i^z$ then $(w, v) \in P_i^z$
S15	if $(w, w') \in P_i^x$ and $(w, v) \in P_i^z$ then $(w', v) \in P_i^z$

It is straightforward to prove that every operator $Bel_i^{\geq x}$ satisfies Axioms 4 and 5. Indeed, the following two formulas are derivable from Axioms **Inc**$_{x,y}$, **PIntrPoss** and **NIntrPoss**, for every $x \in I$ and $i \in AGT$:

- $Bel_i^{\geq x}\varphi \rightarrow Bel_i^{\geq x}Bel_i^{\geq x}\varphi$
- $\neg Bel_i^{\geq x}\varphi \rightarrow Bel_i^{\geq x}\neg Bel_i^{\geq x}\varphi$.

The operators of type $Bel_i^{\geq x}$ introduced above enable to specify a concept of graded belief of the form "agent i has a belief with strength x that φ is true" (or "agent i believes with strength x that φ is true") in which the *exact* strength of the agent's belief is specified. This concept is a fundamental building block for a characterization of the concept of graded trust.

We assume that "agent i has a belief with strength x that φ is true" (or "agent i believes with strength x that φ is true"), noted $Bel_i^x\varphi$, if and only if "agent i believes φ

at least with strength x and, there is no $y > x$ such that agent i believes φ at least with strength y". Formally:

$$Bel_i^x\varphi \stackrel{\text{def}}{=} Bel_i^{\geq x}\varphi \wedge \bigwedge_{y\in I, x<y} \neg Bel_i^{\geq y}\varphi$$

We can prove that every operator Bel_i^x satisfies Axioms 4 and 5. Indeed, the following two formulas are derivable from Axioms $\mathbf{Inc}_{x,y}$, **PIntrPoss** and **NIntrPoss**, for every $x \in I$ and $i \in AGT$:

- $Bel_i^x\varphi \rightarrow Bel_i^x Bel_i^x\varphi$
- $\neg Bel_i^x\varphi \rightarrow Bel_i^x \neg Bel_i^x\varphi$.

6.2 A Definition of Graded Trust

The concept of graded trust we are interested in here is the following.

Definition 8. ***GRADED TRUST IN THE TRUSTEE'S ACTION.*** *Agent i trusts agent j to do α with respect to the achievement of φ with a certain strength x if and only if i has the achievement goal that φ and i believes with strength x that:*

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.*

Definition 8, which is the variant of Definition 1 for graded trust, can be formally translated by means of the operator Bel_i^x:

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

$$AGoal_i\varphi \wedge Bel_i^x(After_{j:\alpha}\varphi \wedge Can_j(\alpha) \wedge Int_j(\alpha))$$

$Trust^x(i, j, \alpha, \varphi)$ is meant to stand for: i trusts j to do α with respect to the achievement of φ with a certain strength x.

Given the logical equivalence between the formula $Can_j(\alpha) \wedge Int_j(\alpha)$ and formula $Does_{j:\alpha}\top$, the definition of graded trust can be simplified as follows:

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

Starting from this concept of graded trust, it is interesting to study how the graded beliefs of the truster about different properties of the trustee are combined in order to assess the trustworthiness of the trustee with respect to a certain task. For instance, suppose that i (the truster) has the achievement goal that φ. Furthermore, i believes with strength x that j (the trustee) has the capacity to do α; i believes with strength y that j, by doing α, will ensure that φ; and i believes with strength z that j intends to do α. Finally, suppose that $z < y < x$. Which is the resulting strength of i's trust in j? From our definition of graded trust, it follows that the strength of i's trust in j is equal to the strength of i's belief with the lowest strength, that is, i's belief that j intends to do α. Indeed, the following formula is derivable in our logic, for every $x, y, z \in I$ such that $z < y < x$:

- $(AGoal_i\varphi \wedge Bel_i^x After_{j:\alpha}\varphi \wedge Bel_i^y Can_j(\alpha) \wedge Bel_i^z Int_j(\alpha)) \rightarrow Trust^z(i, j, \alpha, \varphi)$.

This is a consequence of the more general theorem, derivable from Axiom $\mathbf{Inc}_{x,y}$ and the definition of the operator Bel_i^x, stating that, for every $x_1, x_2, ..., x_m \in I$ such that $x_1 < x_2 < ... < x_m$:

- $(Bel_i^{x_1}\varphi_1 \wedge ... \wedge Bel_i^{x_m}\varphi_m) \rightarrow Bel_i^{x_1}(\varphi_1 \wedge ... \wedge \varphi_m)$.

This means that. under the condition $x_1 < x_2 < ... < x_m$, if agent i believes φ_1 with strength x_1, agent i believes φ_2 with strength x_2,..., and agent i believes φ_m with strength x_m then, agent i believes $\varphi_1 \wedge ... \wedge \varphi_m$ with minimal strength x_1.

It is worth noting that, starting from the previous concept of graded trust, the analysis of information sources given in Sections 4 and 5 can be refined by investigating the relationships between graded trust and graded beliefs about the properties of information sources. For instance, it would be interesting to study variants of Theorem 4 in which formulas expressing graded beliefs about properties of information sources appear in the antecedent and formulas expressing graded trust in the information source appear in the consequent of the implication. We postpone this kind of analysis to future work.

6.3 Discussion: Regularity Levels

The solution to the characterization of graded trust sketched in the previous two sections is aimed at capturing the truster's *uncertainty* in the attribution of certain properties to a given target j (Definition 8). Let us briefly discuss a different solution which consists in specifying the truster's belief that a given property belongs *more or less* to a given target j.

Consider for instance the analysis of the properties of information sources developed in Section 4. All properties of information sources have been expressed in a conditional form and have only been considered as binary properties (e.g. according to the definition of competence either j is competent about P or j is not competent about P). A refinement of the analysis of trust in information sources consists in supposing that the properties of information sources are graded, i.e. the property that an agent i ascribes to the information source j is satisfied with a certain degree h. For example, it may be that agent i believes that the information source j is competent with a certain degree h. In order to capture this graded aspect of the properties of an information source, conditional operators of type $\Rightarrow_h$ should be introduced.

A formula $\varphi \Rightarrow_h \psi$ can be interpreted as the fact that there is no strict regularity in the relationship between the fact φ and the fact ψ, i.e. the set of circumstances where φ holds is *included* with a certain degree h in the set of circumstances where ψ holds.[8] For example, the formula $Bel_j P \Rightarrow_h P$ denotes that the inclusion level of the set of circumstances where $Bel_i P$ holds in the set of circumstances where P holds is h. Now, the fact that i believes that the information source j is competent with a certain degree h is represented by $Bel_i(Bel_j P \Rightarrow_h P)$.

[8] Here the term "regularity" is taken with a similar meaning as A.J.I. Jones does in [21] when he says: "there exists a regularity in y's behavior, so that under particular kinds of circumstances y exhibits a particular kind of behavior". The only difference is that here the word "regularity" is not restricted to behavior, but it can also be used for mental attitudes like competence or sincerity.

We think that the basic principles of conditionals of the form $\Rightarrow_h$ should be compatible with an interpretation of a formula $\varphi \Rightarrow_h \psi$ in terms of a conditional probability. However, we leave open the possibility for other interpretations of this kind of formulas (on this point, see also [10]).

In this probabilistic interpretation the meaning of the regularity level h in $\varphi \Rightarrow_h \psi$ is $h = Pr(\psi|\varphi)$, that is, h is the probability that ψ holds when φ holds.

The most important rule to be defined in the axiomatization of the conditional operators $\Rightarrow_h$ is the rule of detachment. That is, assuming that $\varphi \Rightarrow_h \psi$ holds and φ holds, what can be inferred about ψ? In this paper we do not give an answer to this question. However, it seems reasonable to accept the following rule for detachment $\varphi \Rightarrow_h \psi \rightarrow (\varphi \rightarrow \psi^h)$, after having adopted the notation $\psi^h \stackrel{\text{def}}{=} \top \Rightarrow_h \psi$.

7 Related Works and Conclusion

Although there are several comprehensive logical models of security in the recent literature in AI and computer science where the properties of a communication system such as privacy, confidentiality, availability, integrity, authentication are modeled (e.g. [4]), there is still a pressing need for elaborating more precise and general models of reasoning about trust. Indeed, the specification of trust reasoning is typically in the province of this discipline. Logical models of trust have been focused almost exclusively on informational trust, i.e. trust in information sources [8,9,22,27]. In these logics a certain agent is said to trust another agent if the former agent believes what the other agent says or that the information communicated to him by the other agent is reliable. Some authors have introduced trust in information sources as a primitive concept [8,27] whereas other authors have reduced it to a particular kind of belief of the truster [9,22].

In this work we have provided a more general logical analysis of trust and shown that trust in information sources is only a particular instance of a more general concept of trust. We have modeled the properties of information sources such as sincerity, validity, competence, etc. and shown that some of them are epistemic supports of trust in an information source, that is, they are sufficient conditions for trusting an information source to inform whether a certain fact is true (e.g. Theorem 4).

We have devoted special emphasis to the formalization of the motivational aspect of trust. In our perspective, a logic of trust must enable reasoning about goals of agents. In fact, agent i's trust in agent j necessarily involves a main and primary motivational component which is a goal of the truster. If i trusts j then, necessarily i trusts j because i has some goal and thinks that j has the right properties to ensure that such a goal will be achieved. In this sense, our approach is different from the approach proposed in [21] in which the motivational aspect of trust is ignored, and trust is characterized only in terms of two beliefs of the truster: the truster's belief that a certain rule or regularity applies to the trustee (called "rule belief"), and the truster's belief that the rule or regularity is going to be followed by the trustee (called "conformity belief").

Directions for future research are manifold. First of all, our future future works will be devoted to better study the solution for the specification of graded beliefs and graded trust presented in Section 6. Special emphasis will be devoted to compare our solution with existing logical approaches to graded beliefs (e.g. [12,16]) and graded trust

(e.g. [24]) based on probability theory. Secondly, our future works will be devoted to improve over the formalization of information sources presented in this paper. For instance, in our current definitions of the properties of information sources entailment is formalized by a material implication. In the future we will substitute material implication with some form of conditional which is more adequate for our purpose of formalization.

References

1. Belnap, N., Perloff, M., Xu, M.: Facing the future: agents and choices in our indeterminist world. Oxford University Press, New York (2001)
2. Blackburn, P., de Rijke, M., Venema, Y.: Modal Logic. Cambridge University Press, Cambridge (2001)
3. Bratman, M.: Intentions, plans, and practical reason. Harvard University Press (1987)
4. Burrows, M., Abadi, M., Needham, R.M.: A logic of authentication. ACM Transactions on Computer Systems 8(1), 18–36 (1990)
5. Castelfranchi, C., Falcone, R.: Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In: Proceedings of the Third International Conference on Multiagent Systems (ICMAS 1998), pp. 72–79 (1998)
6. Chellas, B.F.: Modal logic: an introduction. Cambridge University Press, Cambridge (1980)
7. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artificial Intelligence 42, 213–261 (1990)
8. Dastani, M., Herzig, A., Hulstijn, J., van der Torre, L.: Inferring trust. In: Leite, J., Torroni, P. (eds.) CLIMA 2004. LNCS, vol. 3487, pp. 144–160. Springer, Heidelberg (2005)
9. Demolombe, R.: To trust information sources: a proposal for a modal logical framework. In: Castelfranchi, C., Tan, Y.-H. (eds.) Trust and Deception in Virtual Societies. Kluwer, Dordrecht (1999)
10. Demolombe, R.: Reasoning about trust: a formal logical framework. In: Jensen, C., Poslad, S., Dimitrakos, T. (eds.) iTrust 2004. LNCS, vol. 2995, pp. 291–303. Springer, Heidelberg (2004)
11. Demolombe, R., Lorini, E.: A logical account of trust in information sources. In: Proceedings of the Eleventh International Workshop on Trust in Agent Societies (2008)
12. Fagin, R., Halpern, J.: Reasoning about knowledge and probability. Journal of the Association for Computing Machinery 41(2), 340–367 (1994)
13. Fagin, R., Halpern, J., Moses, Y., Vardi, M.: Reasoning about Knowledge. MIT Press, Cambridge (1995)
14. 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)
15. Gerbrandy, J.: Bisimulations on Planet Kripke. PhD thesis, University of Amsterdam, The Netherlands (1999)
16. Halpern, J.Y.: Reasoning about uncertainty. MIT Press, Cambridge (2003)
17. Harel, D., Kozen, D., Tiuryn, J.: Dynamic Logic. MIT Press, Cambridge (2000)
18. Herzig, A., Longin, D.: Sensing and revision in a modal logic of belief and action. In: Proceedings of the 15th European Conference on Artificial Intelligence (ECAI 2002). IOS Press, Amsterdam (2002)
19. Hintikka, J.: Knowledge and Belief. Cornell University Press, New York (1962)
20. Huynh, T.G., Jennings, N.R., Shadbolt, N.R.: An integrated trust and reputation model for open multi-agent systems. Journal of Autonomous Agent and Multi-Agent Systems 13, 119–154 (2006)

21. Jones, A.J.I.: On the concept of trust. Decision Support Systems 33(3), 225–232 (2002)
22. Jones, A.J.I., Firozabadi, B.S.: On the characterization of a trusting agent: Aspects of a formal approach. In: Castelfranchi, C., Tan, Y.H. (eds.) Trust and Deception in Virtual Societies, pp. 55–90. Kluwer, Dordrecht (2001)
23. Jonker, C.M., Treur, J.: Formal analysis of models for the dynamics of trust based on experiences. In: Garijo, F.J., Boman, M. (eds.) MAAMAW 1999. LNCS, vol. 1647, pp. 221–231. Springer, Heidelberg (1999)
24. Jøsang, A.: A logic for uncertain probabilities. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9(3), 279–311 (2001)
25. Laverny, N., Lang, J.: From knowledge-based programs to graded belief-based programs, part ii: off-line reasoning. In: Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005), pp. 497–502. Professional Book Center (2005)
26. Lewis, D.: Counterfactuals. Harvard University Press (1973)
27. Liau, C.J.: Belief, information acquisition, and trust in multi-agent systems: a modal logic formulation. Artificial Intelligence 149, 31–60 (2003)
28. 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)
29. Lorini, E., Herzig, A.: A logic of intention and attempt. Synthese 163(1), 45–77
30. Lorini, E., Herzig, A., Castelfranchi, C.: Introducing "attempt" in a modal logic of intentional action. In: Fisher, M., van der Hoek, W., Konev, B., Lisitsa, A. (eds.) JELIA 2006. LNCS (LNAI), vol. 4160, pp. 280–292. Springer, Heidelberg (2006)
31. Marsh, S.: Formalising Trust as a Computational Concept. PhD thesis, University of Stirling, Scotland (1994)
32. Sabater, J., Sierra, C.: Regret: a reputation model for gregarious societies. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002), pp. 475–482. ACM Press, New York (2001)
33. Scherl, R.B., Levesque, H.: Knowledge, action, and the frame problem. Artificial Intelligence 144, 1–39 (2003)
34. Spohn, W.: Ordinal conditional functions: a dynamic theory of epistemic states. In: Harper, W.L., Skyrms, B. (eds.) Causation in decision, belief change and statistics, pp. 105–134. Kluwer, Dordrecht (1998)
35. van Benthem, J.: Correspondence theory. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, 2nd edn., vol. 3, pp. 325–408. Kluwer Academic Publishers, Dordrecht (2001)
36. Von Wright, G.H.: Norm and Action. Routledge and Kegan, London (1963)
37. Witkowski, M., Artikis, A., Pitt, J.: Experiments in building experiental trust in a society of objective-trust based agents. In: Castelfranchi, C., Tan, Y.H. (eds.) Trust and Deception in Virtual Societies, pp. 111–132. Kluwer Academic Publishers, Dordrecht (2001)