Analytical Decomposition of Trust in Terms of Mental and Social Attitudes

Robert Demolombe

Abstract Trust is defined as a truster's belief in some properties. At the beginning they are to reach a goal and then they are refined in trust in some trustee's property from which the truster can infer that his goal will be reached. This property may be the trustee's ability to bring it about that the goal is reached which can itself be derived from the trustee's intention to reach this goal. Then, we show that this intention may be adopted by the trustee depending on three kinds of social relationships: compliance of norms, mutual commitment with another agent or willingness to act without any compensation.

This analytical decomposition is formalized in a modal logic with a conditional connective. However, the technical details that could prevent an intuitive reading are omitted.

Keywords Trust • Ability • Willingness • Compliance • Modal logic

1 Introduction

Trust can be analyzed to answer three kinds of questions: "what is the definition of what we call trust?", "on what grounds would someone trust in something?" and "for which purpose trust can be used?". Here we concentrate on trust definition and on some types of trust supports.

There are many definitions of the notion of trust (Castelfranchi and Falcone 2001, 2010; Bacharach and Gambetta 2001; Demolombe 2001, 2004; Demolombe and Liau 2001), nevertheless most of them share the idea that trust is a kind of belief about something. In Jones (2002) and Jones and Firozabadi (2001) Andrew J. I. Jones has shown that these beliefs are a rather complex type of beliefs that combines beliefs in the regularity of some property which may have exception and beliefs in the fact that these exceptions will not arise in the current situation. Since it is not the main topic of this work to characterize the kind of belief which is involved in trust

R. Demolombe (✉)
Institut de Recherche en Informatique de Toulouse, Toulouse, France
e-mail: robert.demolombe@orange.fr

A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5,
DOI 10.1007/978-3-319-21732-1_3

definition we have accepted a very crude definition which is formally represented in epistemic logic by a system of type K (see Chellas 1988).

The supports of belief can be classified into the following categories:

1. series of truster's previous experiences which show the regularity of some property,
2. information transmitted by trusted information sources about this regularity (see Demolombe 2001, 2011),
3. analytical decomposition in function of trust in other properties which are themselves supported by grounds of the type 1, 2 or 3.

The decomposition of type 3 allows to logically derive trust in something from trust in other things. The main topic of this paper is to investigate this decomposition. The formalization help to show what are the trustee's properties that are relevant for this decomposition. Roughly speaking they are mental attitudes or social attitudes of the agents. The formalization in modal logic of these attitudes is only motivated by the objective to propose as far as possible clear definitions of these attitudes and of their relationships. However, we shall not try to give formal detailed definitions of notions like "intention to do" or "attempt to do" which are quite complex and rather controversial. On the contrary we have preferred to leave open these refinements when they have no influence on the decomposition. According to this approach limited information is given about the formal properties of the logic which is presented in the annex.

In the next section, after the informal definitions of the logical framework, we present the starting point of the decomposition and we split it into two categories: trust in the possibility to reach a state of affairs and trust in the possibility to maintain a state of affairs. The decompositions based on these two categories are respectively analyzed in Sects. 3 and 4. In Sect. 5 is presented a comparison with other works and the last section summarizes our conclusions. In the annex some details are given about formal properties.

2 Initial Trust Definition

The formal language that will be used in the rest of the paper is defined as follows: *ATOM*: set of atomic propositions, *AGENT*: set of agents, *MODAL*: set of modal operators.

The language is the set of formulas defined by the following BNF:

$$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid \phi \Rightarrow \phi \mid M_i\phi$$

where p ranges over *ATOM*, i range over *AGENT* and M ranges over *MODAL*.

The intuitive meaning of the logical connective $\phi \Rightarrow \psi$ is: ϕ entails ψ.

The set of modal operators *MODAL* and their intuitive meaning is:

$Bel_i\phi$: i believes that ϕ holds.
$Goal_i\phi$: i's goal is that ϕ holds.

$\Box\phi$: ϕ holds now and it will hold always in the future.
$\Diamond\phi$: there is a future instant where ϕ holds (the operator $\Diamond$ is defined from $\Box$ by: $\Diamond\phi \stackrel{\text{def}}{=} \neg\Box\neg\phi$).
$Attempt_i\phi$: i attempts to bring it about that ϕ.[1]
$Int_i\phi$: i's intention is that ϕ holds.
$Obg_j\phi$: it is obligatory that j brings it about that ϕ.
$Ask_{i,j}\phi$: i asks j to bring it about that ϕ.
$Commit_{j,i}\phi$: j commits himself with regard to i to bring it about that ϕ.

In the presented analysis it is assumed that initial trust has practical motivations. That is, trust is a truster's belief in the fact that if he has some given goal, then this goal will be reached. The conditional form of this belief shows that what is trusted is not a property that holds just in the present situation, but rather it holds in every situations where the truster wants to reach this given goal. That is why trust is formally represented by a conditional operator instead of material implication and it has the general form:

$$Bel_i(Goal \Rightarrow GoalReached)$$

where i denotes the truster.

Nevertheless, in most of the real situations this set of situations is restricted to some particular context. For instance, if the truster trusts in the fact that if he wants to take a taxi, then he will find a taxi, his trust may be restricted not to be after midnight and to stay close to downtown. This restriction could be formally represented by the formula:

$$Bel_i(context \Rightarrow (Goal \Rightarrow GoalReached))$$

However, to avoid overly complex formula in the following the context will remain implicit.

This initial trust definition is refined depending on the type of goal. We have considered a first type of goal which is **to reach** a state of affairs. If this state of affairs is denoted by the formula ϕ, the antecedent of the conditional property is: $\neg\phi \wedge Goal_i\Diamond\phi$, which means that ϕ does not hold in the present situation and i's goal is that it holds in the future, and the consequent is: $\Diamond\phi$, which means that the goal ϕ will be reached at some future instant. Then, this type of trust is represented by:

(R1) $\quad Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$

In addition, i's goal is not just that ϕ holds at some instant in the future, rather, this instant should happen before a given deadline. For instance, if the truster wants to take a taxi, he is expecting that the taxi will come before some delay. Also, he is aware of the fact that what he trusts may change in a long term future. Then, a more realistic definition would take the form:

[1]The meaning of the operator "to bring it about that ϕ" can be found in Pörn (1977).

$$Bel_i(Until(d, (\neg\phi \wedge Goal_i(Before(d', \phi)) \Rightarrow Before(d', \phi))))$$

where $Until(d, \psi)$ means that ψ will hold until instant d and $Before(d', \phi)$ means that ϕ will hold before the instant d'.

However, in the following these temporal refinements will be ignored and we use definition (R1) to avoid overly complex formulas whose intuitive understanding is not easy. The same approach is adopted throughout the rest of the paper.

Examples of (R1).

- i believes that if his car is out of order ($\neg\phi$) and his goal is to have his car repaired ($Goal_i\Diamond\phi$), then it will be repaired ($\Diamond\phi$).
- i believes that if he has a flu and his goal is to be cured, then he will be cured.

Notice that in these examples there is no explicit reference to some trustee.

The second type of goal is **to maintain** a state of affairs ϕ. Here, the antecedent is denoted by: $\phi \wedge Goal_i\Box\phi$, which means that ϕ holds in the present situation and i's goal is that ϕ will hold for ever, and the consequent is: $\Box\phi$, which means that ϕ will hold for ever. Then, this type of trust is represented by:

(M1) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

Example of (M1). Let's assume that i is visiting a dangerous city. i believes that if he is alive (ϕ) and his goal is to stay alive ($Goal_i\Box\phi$), then he will stay alive ($\Box\phi$).

In the next sections we analyze from which kinds of trusts (R1) and (M1) can be derived.

3 To Reach a State of Affairs

Trust of the type (R1) may be derived from the fact that there exists some agent j such that i believes that if he has the goal to reach a state where ϕ holds, then j will attempts to bring it about that ϕ (which is represented by (R2)) and i also believes that if j attempts to bring it about that ϕ, then ϕ will hold (represented by (S2)).

(R2) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi)$
(S2) $Bel_i(Attempt_j\phi \Rightarrow \Diamond\phi)$

Both (R2) and (S2) are new kinds of trust. The intuitive meaning of (S2) is that i trusts j in his ability to bring it about that ϕ. If ability is defined as follows:

$$Able_j\phi \stackrel{\text{def}}{=} Attempt_j\phi \Rightarrow \Diamond\phi$$

(S2) can be represented by: $Bel_i(Able_j\phi)$.

It can be easily shown (see Property RS2 in the Annex) that (R2) and (S2) entail (R1) and this shows that they are a possible analytical decomposition of (R1).

Examples of (R2) and (S2). There exists some agent j such that i believes that if his car is out of order and his goal is to have his car repaired ($\neg\phi \wedge Goal_i\Diamond\phi$), then

j will attempt to repair his car ($Attempt_j\phi$) and i believes that j is able to repair his car in the sense that if j attempts to repair his car, then his car will be repaired.

Trust of the type (R2) may itself be derived from the fact that i believes that if he has the goal to reach the state ϕ, then j will adopt the intention to bring it about that ϕ (represented by (R3)) and i also believes that if j has the intention to bring it about that ϕ, then j will attempt to bring it about that ϕ. That is represented by (S3)).

(R3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_j\phi)$
(S3) $Bel_i(Int_j\phi \Rightarrow Attempt_j\phi)$

We use the word "determined" to speak about the j's property represented by (S3). This determination property may seem to be obvious, however there are irresolute or indecisive agents who may have the intention to bring it about that ϕ and never start to act. This property is formally defined by:

$$Determined_j\phi \stackrel{\text{def}}{=} Int_j\phi \Rightarrow Attempt_j\phi$$

and trust (S3) is represented by: $Bel_i(Determined_j\phi)$.

It can be shown that (R3) and (S3) entail (R2) (see Property RS23 in the Annex).

Examples of (R3) and (S3). i believes that if his car is out of order and his goal is to have his car repaired, then j will adopt the intention to repair his car ($Int_j\phi$) and i believes that j is determined to repair i's car ($Determined_j\phi$) in the sense that if he (j) has the intention to repair his car, he will attempt to repair it.

Trust of the type (R3) can be derived from several different kinds of trust. Each one is a possible answer to i's question: what could be a justification of the fact that j has adopted the intention to satisfy i's goal?

There are 3 basic answers to this question[2]:

1. j is obliged to bring it about that ϕ
2. if j brings it about that ϕ, then i will give to j some compensation
3. j is willing to help i without any compensation

Case 1. In case 1 trust (R3) can be derived from the fact that i believes that if he has the goal to reach the state ϕ, then j will believe that he (j) is obliged to bring it about that ϕ (represented by (R4.1)) and i also believes that if j believes that he is obliged to bring it about that ϕ, then j will adopt the intention to bring it about that ϕ (represented by (S4.1)).

(R4.1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Bel_jObg_j\phi)$
(S4.1) $Bel_i(Bel_jObg_j\phi \Rightarrow Int_j\phi)$

It is worth noting that if j ignores that he is obliged to bring it about that ϕ, there is no chance that this obligation influences j's attitude. That is why in (R4.1) and (S4.1) we have $Bel_jObg_j\phi$ instead of $Obg_j\phi$.

[2]We do not pretend that these three possibilities are exhaustive but we think that they cover most of the situations.

Both (R4.1) and (S4.1) are new kinds of trust. The intuitive meaning of (S4.1) is that i trusts j in his compliance with the obligation to bring it about that ϕ. If that type of compliance is defined as follows:

$$CompObg_j\phi \stackrel{\text{def}}{=} Bel_jObg_j\phi \Rightarrow Int_j\phi$$

(S4.1) can be represented by: $Bel_i(CompObg_j\phi)$.

It can be shown that (R4.1) and (S4.1) entail (R3).

Examples of (R4.1) and (S4.1). Here, it is assumed that j is a car mechanic and i is an ambulance driver, and there is a norm which says that if an ambulance is out of order, car mechanics are obliged to repair the ambulance. In this context i believes that if his ambulance is out of order and his goal is to have his ambulance repaired, then j will believe that it is obligatory that he repairs i's ambulance ($Bel_jObg_j\phi$) and i believes that j ordinarily complies with obligation ($CompObg_j\phi$) in the sense that if j believes that it is obligatory that he repairs i's ambulance, then j will adopt the intention to repair it.

Trust (R4.1) can itself be derived from the fact that i believes that if he has the goal to reach the state ϕ, then there is some agent k who will ask j to bring it about that ϕ (represented by (R5.1)) and i also believes that if k asks j to bring it about that ϕ, then j will believe that he is obliged to bring it about that ϕ (represented by (S5.1)).

(R5.1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Ask_{k,j}\phi)$

(S5.1) $Bel_i(Ask_{k,j}\phi \Rightarrow Bel_jObg_j\phi)$

The intuitive meaning of (S5.1) is that i believes if k order j to bring it about that ϕ, then j will believe that k has ordered him to bring it about that ϕ and i also believes that k has authority (in the sense of "has institutional power" Jones and Sergot 1996) to order j to bring it about that ϕ. Of course, it is not excluded that k was i himself. If that type of authority is defined as follows:

$$Authorized_{k,j}\phi \stackrel{\text{def}}{=} Ask_{k,j}\phi \Rightarrow Obg_j\phi$$

(S5.1) can be derived from: $Bel_i(Ask_{k,j}\phi \Rightarrow Bel_j(Ask_{k,j}\phi))$ and $Bel_i(Bel_j(Authorized_{k,j}\phi))$.

It can be shown that (R5.1) and (S5.1) entail (R4.1).

Examples of (R5.1) and (S5.1). Now, it is assumed that there exists a policeman k who has authority to order to the car mechanic j to repair i's ambulance ($Authorized_{k,j}\phi$). In this context i believes that if his ambulance is out of order and his goal is to have his ambulance repaired, then k will ask j to repair it ($Ask_{k,j}\phi$) and i believes that if k asks j to repair i's ambulance, then j will believe that it is obligatory that he repairs it ($Bel_jObg_j\phi$).

Case 2. In case 2 (R3) can be derived from the fact that i believes that if he has the goal to reach the state ϕ, then j will commit with respect to i to bring it about that ϕ and i will commit with respect to j to bring it about that ψ (represented

by (R4.2)) and i also believes that if this mutual commitment holds, then j will adopt the intention to bring it about that ϕ (represented by (S4.2)).

(R4.2) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow MutualCommit_{j,i}(\phi, \psi))$
(S4.2) $Bel_i(MutualCommit_{j,i}(\phi, \psi) \Rightarrow Int_j\phi)$

where $MutualCommit_{j,i}(\phi, \psi)$ is defined by:

$$MutualCommit_{j,i}(\phi, \psi) \stackrel{\text{def}}{=} (Commit_{j,i}\phi) \wedge (Commit_{i,j}\psi)$$

The intuitive meaning of (S4.2) is that if there is a mutual commitment between j and i, then j will comply his commitment.

If this compliance is formally defined by:

$$CompCommit_{j,i}(\phi, \psi) \stackrel{\text{def}}{=} MutualCommit_{j,i}(\phi, \psi) \Rightarrow Int_j\phi$$

(S4.2) can be represented by: $Bel_i(CompCommit_{j,i}(\phi, \psi))$.

It can be shown that (R4.2) and (S4.2) entail (R3).

Examples of (R4.2) and (S4.2). Here, it is no more assumed that i is an ambulance driver. i believes that if his car is out of order and his goal is to have his car repaired, then j will commit himself to repair the car ($Commit_{j,i}\phi$) and i will commit himself to pay j ($Commit_{i,j}\psi$) and i believes that if this mutual commitment between i and j ($MutualCommit_{j,i}(\phi, \psi)$) holds, then j will adopt the intention to repair his car ($Int_j\phi$).

Case 3. In case 3 (R3) can be derived from the fact that i believes that if he has the goal to reach the state ϕ, then j will be aware of his goal (represented by (R4.3)) and i also believes that if j is aware of i's goal, then j will adopt the intention to bring it about that ϕ (represented by (S4.3)).

(R4.3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Bel_jGoal_i\Diamond\phi)$
(S4.3) $Bel_i(Bel_jGoal_i\Diamond\phi \Rightarrow Int_j\phi)$

The intuitive meaning of (S4.3) is that j is willing to satisfy i's goal without any compensation; j's attitude could also be defined as altruist. If j's willingness is defined as follows:

$$Willing_{j,i}\phi \stackrel{\text{def}}{=} Bel_jGoal_i\Diamond\phi \Rightarrow Int_j\phi$$

(S4.3) can be represented by: $Bel_i(Willing_{j,i}\phi)$.

It can be shown that (R4.3) and (S4.3) entail (R3).

Examples of (R4.3) and (S4.3). Let's assume now that j can observe that i's car is out of order. i believes that if his car is out of order and his goal is to have his car repaired, then j will believe that i's goal is to have his car repaired and i also believes that if j will believe that i's goal is to have his car repaired, then j will adopt the intention to repair his car.

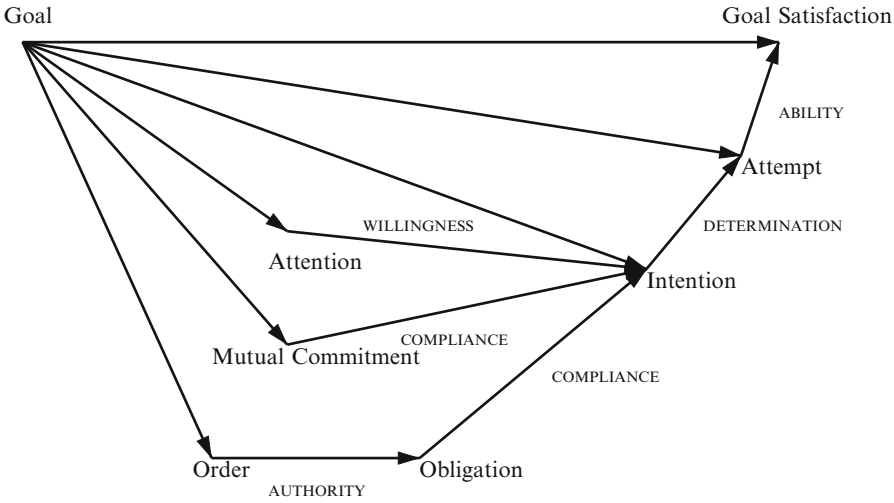

Fig. 1 Trust analytical decomposition

Figure 1 gives a global picture of the different types of trust and of their relationships.

In the previous analysis of the different types of trust it has implicitly been assumed that the truster i knows who is the trustee j and if i trusts j with respect to several properties, for instance: ability and determination, these properties can be represented by the formulas:

(S2) $Bel_i(Able_j\phi)$
(S3) $Bel_i(Determined_j\phi)$
(R3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_j\phi)$

Due to the logical properties of modality *Bel*, the following can be inferred:

$$Bel_i((Determined_j\phi) \wedge (Able_j\phi) \wedge (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_j\phi))$$

which entails:

(ExR3) $\exists x Bel_i((Determined_x\phi) \wedge (Able_x\phi) \wedge (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_x\phi))$

In that formula the variable x which denotes the trustee is interpreted **de re**, that is, the truster i knows who is x. However, there may be situations where i believes that there exists some trustee x who holds the properties represented by: $(Determined_x\phi) \wedge (Able_x\phi) \wedge (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_x\phi)$ even if i does not know such an x. In these situations the existential variable x has to be interpreted **de dicto** in the formula:

(ExR'3) $Bel_i\exists x((Determined_x\phi) \wedge (Able_x\phi) \wedge (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_x\phi))$

It can be shown that both (ExR3) and (ExR'3) entail (R1).

4 To Maintain a State of Affairs

If the truster's goal is to maintain a state of affairs we can follow a very similar approach to analyze analytical trust decomposition as in the case where his goal is to reach a state of affairs. However, there are some significant differences.

The decomposition of the initial trust:

(M1) $Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

in terms of trustee's ability requires two assumptions.

The first one is that no agent who is able to bring it about that $\neg\phi$ will attempt to bring it about that $\neg\phi$ if i's goal is that ϕ does not change. The second one (we call it "persistence assumption") is that if the first assumption is satisfied, then i believes that if his goal is to maintain the state of ϕ, then ϕ remain unchanged.

These properties can be formally represented by:

(M2) $\forall x Bel_i(Able_x\neg\phi \rightarrow (\phi \wedge Goal_i\Box\phi \Rightarrow \neg Attempt_x\neg\phi))$

(N2) $\forall x Bel_i(Able_x\neg\phi \rightarrow (\phi \wedge Goal_i\Box\phi \Rightarrow \neg Attempt_x\neg\phi)) \rightarrow Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \Box\phi)$

It is obvious that (M2) and (N2) entail (M1).

Notice that $(Able_x\neg\phi) \wedge (Attempt_x\neg\phi)$ entails $\Diamond\neg\phi$ which will mean that i's goal $Goal_i\Box\phi$ fails (see Property MN2 in the Annex). That is why in (M2) it is required that $(\phi \wedge Goal_i\Box\phi \Rightarrow \neg Attempt_x\neg\phi)$ holds.

Another tempting formulation of what is represented by (M2) could be: there is no x such that i believes that x is able to bring it about that $\neg\phi$ and x attempts to bring it about that $\neg\phi$ when i's goal is that ϕ does not change:

(M2bis) $\neg\exists x Bel_i(Able_x\neg\phi \wedge (\phi \wedge Goal_i\Box\phi \Rightarrow Attempt_x\neg\phi))$

However, (M2bis) is logically equivalent to:

(M2ter) $\forall x < Bel_i > \neg(Able_x\neg\phi \wedge (\phi \wedge Goal_i\Box\phi \Rightarrow Attempt_x\neg\phi))$

where $< Bel_i >$ is an abbreviation for the possibility operator $\neg Bel_i\neg$, and (M2ter) is consistent with:

(M2qrt) $\forall x < Bel_i > (Able_x\neg\phi \wedge (\phi \wedge Goal_i\Box\phi \Rightarrow Attempt_x\neg\phi))$

which means that it is consistent with what i believes that agents who are able to bring it about that $\neg\phi$ attempt to bring it about that $\neg\phi$ in circumstances where i's goal is that ϕ remains unchanged. It is clear that in this situation i cannot trust in the fact that ϕ will remain unchanged and that (M2bis) must be rejected.

Another wrong variant of (M2) is:

(M2qnt) $\forall x(Able_x\neg\phi \rightarrow Bel_i(\phi \wedge Goal_i\Box\phi \Rightarrow \neg Attempt_x\neg\phi))$

This formalization is wrong because in (M2qnt) i knows who agents x are but he does not know that these x are able to bring it about that $\neg\phi$. Therefore, i does not know that the set of x who do not attempt to bring it about that $\neg\phi$ contains all the agents who are able to bring it about that $\neg\phi$. That is why (M2qnt) must also be rejected.

In (M2) and (N2) the universally quantified formula x is interpreted *de re*, if it is interpreted *de dicto* we have:

(M'2) $Bel_i \forall x(Able_x \neg\phi \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Attempt_x \neg\phi))$
(N'2) $Bel_i \forall x(Able_x \neg\phi \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Attempt_x \neg\phi)) \rightarrow$
$Bel_i(\phi \wedge Goal_i \Box\phi \Rightarrow \Box\phi)$

Since the two interpretations are very close from a formal point of view, in the following we concentrate only on the *de dicto* interpretation.

Examples of (M'2) and (N'2). In the same context as in the example of (M1), i believes that for every x who is able to kill him ($Able_x \neg\phi$), if i's goal is to stay alive ($\phi \wedge Goal_i \Box\phi$), then x will not attempt to kill him ($\neg Attempt_x \neg\phi$). In addition, i believes that in this situation he will stay alive (N'2). In that example the persistence assumption is quite strong since it excludes situations where i could be killed by accident by someone who is not able to kill him, in the sense that if he attempts to kill i he may kill i but that is not guaranteed.

If the truster i believes that the agents who are determined and able to bring it about that $\neg\phi$ do not adopt the intention to bring it about that $\neg\phi$ when i's goal is that ϕ remains unchanged, then i believes that ϕ will remain unchanged (see Property MN3 in the Annex). This situation is formally represented by:

(M3) $Bel_i \forall x((Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Int_x \neg\phi))$
(N3) $Bel_i \forall x((Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Int_x \neg\phi)) \rightarrow$
$Bel_i(\phi \wedge Goal_i \Box\phi \Rightarrow \Box\phi)$

It is clear that (M3) and (N3) entail (M1).

Examples of (M3) and (N3). The example of (M'2) and (N'2) can be extended here. The only difference is that agents x who are determined to kill i do not adopt the intention to kill him.

It is interesting to observe the formal duality between (M3) and (ExR'3):

(ExR'3) $Bel_i \exists x((Determined_x \phi) \wedge (Able_x \phi) \wedge (\neg\phi \wedge Goal_i \Diamond\phi \Rightarrow Int_x \phi))$

According to this duality, for the decomposition corresponding to **case 1** we have:

(M4.1) $Bel_i \forall x((CompObg_x \neg\phi) \wedge (Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow$
$(\phi \wedge Goal_i \Box\phi \Rightarrow \neg Bel_x Obg_x \neg\phi))$
(N4.1) $Bel_i \forall x((CompObg_x \neg\phi) \wedge (Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow$
$(\phi \wedge Goal_i \Box\phi \Rightarrow \neg Bel_x Obg_x \neg\phi)) \rightarrow Bel_i(\phi \wedge Goal_i \Box\phi \Rightarrow \Box\phi)$

Examples of (M4.1) and (N4.1). Like in the examples of (M3) and (N3) it is assumed that there is a criminal organization which can oblige its members to kill somebody. Here, i believes that for every x who complies with the obligations of this organization, if i's goal is to stay alive, then x does not believe that he is obliged to kill i.

If the obligations are analyzed as the results of orders given by authorized agents, we have:

(M5.1) $Bel_i \forall x \forall y((Authorized_{y,x} \neg\phi) \wedge (CompObg_x \neg\phi) \wedge (Determined_x \neg\phi) \wedge$
$(Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Ask_{y,x} \neg\phi))$

(N5.1) $Bel_i \forall x \forall y((Authorized_{y,x} \neg\phi) \wedge (CompObg_x \neg\phi) \wedge (Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Ask_{y,x} \neg\phi)) \rightarrow Bel_i(\phi \wedge Goal_i \Box\phi \Rightarrow \Box\phi)$

If i knows who are the authorized agents, instead of (M5.1) which has the form: (M5.1)$Bel_i \forall x \forall y((Authorized_{y,x} \neg\phi) \ldots)$ we have: $\forall y Bel_i \forall x((Authorized_{y,x} \neg\phi) \ldots)$.

Examples of (M5.1) and (N5.1). Like in the previous example, it can be assumed that there are agents y who the authorized agents are in this organization to create the obligation to kill i by asking some x to kill i, and these agents do not ask to kill i.

For a decomposition corresponding to the **case 2** we have:

(M4.2) $Bel_i \forall x((CompCommit_{x,i}(\neg\phi, \psi) \wedge (Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg MutualCommit_{x,i}(\neg\phi, \psi))$

(N4.2) $Bel_i \forall x((CompCommit_{x,i}(\neg\phi, \psi) \wedge (Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg MutualCommit_{x,i}(\neg\phi, \psi)) \rightarrow Bel_i(\phi \wedge Goal_i \Box\phi \Rightarrow \Box\phi)$

Here i's trust is justified by the fact that there is no mutual commitment between x and i to bring it about that $\neg\phi$.

Examples of (M4.2) and (N4.2). Let's consider now a situation where i is a regular customer of a given hotel. Some days he wants to sleep in the morning (ϕ) and some other days he wants to be woken up ($\neg\phi$). In this context it may be that if i has a mutual commitment with an employee x of the hotel to be woken up and to give him a tip (ψ) in compensation ($MutualCommit_{x,i}(\neg\phi, \psi)$), then x will adopt the intention to wake up i ($CompCommit_{x,i}(\neg\phi, \psi)$). Then, i believes that if x complies with this mutual commitment, if i's goal is not to be woken up, then there will be no such mutual commitment.

For a decomposition corresponding to the **case 3** we have:

(M4.3) $Bel_i \forall x((Willing_{x,i} \neg\phi) \wedge (Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Bel_x Goal_i \Diamond\neg\phi))$

(N4.3) $Bel_i \forall x((Willing_{x,i}(\neg\phi) \wedge (Determined_x \neg\phi) \wedge (Able_x \neg\phi) \rightarrow (\phi \wedge Goal_i \Box\phi \Rightarrow \neg Bel_x Goal_i \Diamond\neg\phi)) \rightarrow Bel_i(\phi \wedge Goal_i \Box\phi \Rightarrow \Box\phi)$

Here, i's trust is justified by the fact that the agents x who are willing and able to bring it about that $\neg\phi$ do not believe that i's goal is to change the status of ϕ.

Examples of (M4.3) and (N4.3). In the same example as for (M4.2) and (N4.2), let's assume that instead of agents x who adopt the intention to wake up i in return for a tip we have agents x whose intention to wake up i is only motivated by the fact that they believe that i's goal is to be woken up ($Willing_{x,i} \neg\phi$). In this context, if i's goal is to sleep, x does not believe that his goal is to be woken up ($\neg Bel_x(Goal_i \Diamond\neg\phi)$) and consequently x does not adopt the intention to wake up i.

5 Comparison with Other Works

At the beginning it was mentioned that we have adopted an extremely crude notion of belief though beliefs play a quite significant role in the notion of trust. In Jones (2002) and Jones and Firozabadi (2001) Andrew J. I. Jones makes the distinction between two kinds of beliefs involved in trust definition: "rule belief"

and "conformity belief". Rule belief expresses a regularity between some state of affairs, formally represented by *context*, and trustee's behavior (more precisely the fact that the trustee brings it about that ϕ, represented by $E_j\phi$). This regularity is represented by [3]:

$$Bel_i(context \rightarrow\!\!\rightarrow E_j\phi)$$

where $\rightarrow\!\!\rightarrow$ is intended to represent a conditional that tolerates exceptions. Conformity belief expresses that exceptional circumstances will not arise on the occasion concerned.

In Demolombe (2009) Demolombe has proposed a notion of graded trust where a distinction is made between the level of uncertainty g of the truster's belief, on the one hand, and the regularity level h of the conditional, on the other hand. That is formally represented by:

$$Bel_i^g(\phi \Rightarrow^h \psi)$$

For instance, in the case of trust in determination we could have:

$$Bel_i^g(Int_j\phi \Rightarrow^h Attempt_j\phi)$$

The relationships between the notions of belief presented above deserve further researches.

The idea of trust decomposition has been introduced by Demolombe in Demolombe (2001) for trust in trustee's epistemic properties. For instance, a "valid" information source is defined as an information source j such that, if j informs the truster i about proposition ϕ, then ϕ holds. This trustee's property is refined in terms of "sincerity" and "competence", where agent j is sincere iff if j informs i about ϕ, then j believes that ϕ holds and agent j is "competent" iff if j believes ϕ, then ϕ holds. Then, if i trusts j in his sincerity and his competence, i can infer that he can trust j in his validity. However, in this work there is no reference to i's goal nor to j's intention.

In Castelfranchi and Falcone (2001, 2010) Castelfranchi et al. assume that the truster i has a goal which is to reach a given situation where ϕ holds and there exists some other agent j, the trustee, such that the truster believes that j can do an action α which has the effect ϕ and j has the intention to do this action. This definition is informally characterized by:

- truster's goal is to reach a situation where the proposition ϕ holds
- the action α has the effect that ϕ holds
- the trustee has the ability and opportunity to do the action α

[3]The notations have been changed in order to make easier the comparison with the presented approach.

- the trustee has the intention to do α

A common feature with the presented approach is that trust definition refers to the truster's goal and also to the trustee's ability and intention to reach a state of affairs. However, there are significant differences. The first one is that situations where the truster's goal is to maintain a state of affairs are ignored. The second one is that there is no refinement of an initial definition in terms of other kinds of trust. For instance, there is no attempt to investigate what could justify the fact that the truster has adopted the intention to do action α.

This approach has been expressed by Lorini and Demolombe in modal logic in Lorini and Demolombe (2008) with some significant improvements. In particular a notion of obedient agent was introduced which is close to what we have called compliance with obligations and the notion of willingness which is rather close to the kind of willingness we have presented above. It is formally defined by[4]:

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

which can be rephrased as: j's goal is that if he believes that i's goal is that j does action α, then j adopts the intention to do α.

From this notion of willingness is defined "positive trust about willingness" as follows:

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

which can be rephrased as: i's goal is that at the next step ϕ holds ($X\phi$) and i believes that, after performance of α, ϕ holds ($After_{j:\alpha}\phi$) and j can do α and j is willing to do α.

Here, the difference with our approach is that the only property assigned to the trustee which has a conditional form is his willingness. The other conditions refer to the current situation.

Another difference is that it is implicitly assumed that if j has the intention to do α, then he does α and if he does α, then ϕ will hold. That is, it is implicitly assumed that j is Determined and Able to do α in the sense we have defined. Also, there is no inclusion of the fact that the motivation to adopt the intention to do α may be that there is a mutual commitment between the truster and the trustee.

In this paper is also defined the notion of "negative trust about willingness" which is formally defined by:

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

[4]We have simplified the formal definition. In the complete definition there is an additional condition which has been introduced in order to avoid some paradoxes due to material implication.

This type of trust has some common features with trust in maintenance of a state of affairs. The difference is that it "guarantees" that j will not prevent ϕ from obtaining, but, it may be that i also believes that another agent than j may prevent ϕ from obtaining.

For instance, in the example of the agent who is in a dangerous city and wants to stay alive, the fact the truster believes that agent j will not adopt the intention to kill him "guarantees" that he will not be killed by this agent but that does not "guarantee" that he will not be killed by another agent.

6 Conclusion

We have shown how the notion of trust in some property can be grounded on trust in other properties. Truster's goal to reach a state of affairs or to maintain a state of affairs can be grounded on trust in trustee's ability to bring it about this state of affairs which can be itself grounded on trustee's determination to attempt to do what he intends to do. The trustees's intention may be grounded itself on his compliance of obligations or by mutual interest and commitment with the truster or by willingness to satisfy truster's goal.

This decomposition has been formalized in conditional logic and in modal logic and we have tried to limit as far as possible the technical details of these logics. In particular, we have adopted a very simple notion of truster's belief which could be refined in the directions mentioned in the comparison with other works. Another possible improvement could be to go further into the analysis of the temporal dimension, in particular the analysis of how trust changes or persists after the truster has used trust to take decisions and after observation of the effects of these decisions.

Acknowledgements I am very grateful to Andrew J.I. Jones for his valuable comments and for his help in the writing of the paper.

Annex

The axiomatics, in addition to the axiomatics of classical propositional calculus, is defined as follows.

The modal operators Bel_i and $\Box$ obey the axiomatics of a normal modal logic of system K.

For the conditional operator we have the following axiom schemas and inference rules:

(EQUIV) If $\vdash \phi \leftrightarrow \phi'$ and $\vdash \psi \leftrightarrow \psi'$, then $\vdash (\phi \Rightarrow \psi) \rightarrow (\phi' \Rightarrow \psi')$
(TRANS) $(\phi_1 \Rightarrow \phi_2) \wedge (\phi_2 \Rightarrow \phi_3) \rightarrow (\phi_1 \Rightarrow \phi_3)$
(DIST) $(\phi_1 \Rightarrow \phi_2) \rightarrow (\phi_1 \rightarrow \phi_2)$

Property RS2.
We have: (R2) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi)$ and (S2) $Bel_i(Attempt_j\phi \Rightarrow \Diamond\phi)$ entail (R1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$.

Proof. From the properties of a system K, from (R2) and (S2) we have:

(1) $Bel_i((\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi) \wedge (Attempt_j\phi \Rightarrow \Diamond\phi))$
From (TRANS), we have :
(2) $(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi) \wedge (Attempt_j\phi \Rightarrow \Diamond\phi) \rightarrow (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$
From Necessitation applied to Bel_i and (2) we have:
(3) $Bel_i((\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi) \wedge (Attempt_j\phi \Rightarrow \Diamond\phi) \rightarrow (\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi))$
From K and (1) and (3) we have:

(R1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$

Property RS23.
We have: (R3) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Int_j\phi)$,
(S3) $Bel_i(Int_j\phi \Rightarrow Attempt_j\phi)$ and (S2) $Bel_i(Attempt_j\phi \Rightarrow \Diamond\phi)$ entail (R1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$.

Proof. With the same kind of proof as for Property RS2, from (R3) and (S3) we have:

(1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow Attempt_j\phi)$
With the same kind of proof, from (1) and (S2) we have:

(R1) $Bel_i(\neg\phi \wedge Goal_i\Diamond\phi \Rightarrow \Diamond\phi)$.

Property MN2.
We have the logical theorem: $Bel_i((Able_x\neg\phi) \wedge (Attempt_x\neg\phi) \rightarrow \Diamond\neg\phi)$.

Proof. From (DIST) and *Able* definition we have:

(1) $(Able_x\neg\phi) \rightarrow (Attempt_x\neg\phi) \rightarrow \Diamond\neg\phi)$
Therefore, we have:
(2) $(Able_x\neg\phi) \wedge (Attempt_x\neg\phi) \rightarrow \Diamond\neg\phi)$

Since Bel_i obeys a system K from (2) we have:

$$Bel_i((Able_x\neg\phi) \wedge (Attempt_x\neg\phi) \rightarrow \Diamond\neg\phi).$$

Property MN3.
We have the logical theorem: $Bel_i((Determined_x\neg\phi) \wedge (Able_x\neg\phi) \wedge (Int_x\neg\phi) \rightarrow \Diamond\neg\phi)$.

Proof. From *Determined* and *Able* definitions, $(Determined_x\neg\phi) \wedge (Able_x\neg\phi)$ is an abbreviation for:

(1) $(Int_x\neg\phi \Rightarrow Attempt_x\neg\phi) \wedge (Attempt_x\neg\phi \Rightarrow \Diamond\neg\phi)$
From (TRANS), (1) entails:

(2) $(Determined_x\neg\phi) \wedge (Able_x\neg\phi) \rightarrow (Int_x\neg\phi \Rightarrow \Diamond\neg\phi)$
From (2) and (DIST) we have:
(3) $(Determined_x\neg\phi) \wedge (Able_x\neg\phi) \rightarrow (Int_x\neg\phi \rightarrow \Diamond\neg\phi)$
And from classical logic (3) entails:
(4) $(Determined_x\neg\phi) \wedge (Able_x\neg\phi) \wedge (Int_x\neg\phi) \rightarrow \Diamond\neg\phi$
Since Bel_i obeys a system K, from (4) we have:
(5) $Bel_i((Determined_x\neg\phi) \wedge (Able_x\neg\phi) \wedge (Int_x\neg\phi) \rightarrow \Diamond\neg\phi)$

References

Bacharach, M., and D. Gambetta. 2001. Trust as type detection. In *Trust and deception in virtual societies*, ed. C. Castelfranchi and Y.-H. Tan. Dordrecht/Boston: Kluwer Academic.

Castelfranchi, C., and R. Falcone. 2001. Social trust: A cognitive approach. In *Trust and deception in virtual societies*, ed. C. Castelfranchi and Y.-H. Tan. Dordrecht/Boston: Kluwer Academic.

Castelfranchi, C., and R. Falcone. 2010. *Trust theory: A socio-cognitive and computational model*. Chichester: Wiley.

Chellas, B.F. 1988. *Modal logic: An introduction*. Cambrige: Cambridge University Press.

Demolombe, R. 2001. To trust information sources: A proposal for a modal logical framework. In *Trust and deception in virtual societies*, ed. C. Castelfranchi and Y.-H. Tan. Dordrecht/Boston: Kluwer Academic.

Demolombe, R. 2004. Reasoning about trust: A formal logical framework. In *Trust management: Second international conference iTrust*, LNCS 2995, ed. C. Jensen, S. Poslad, and T. Dimitrakos. Berlin/London: Springer.

Demolombe, R. 2009. Graded trust. In *Proceedings of the trust in agent societies workshop at AAMAS 2009*, Budapest, ed. R. Falcone, S. Barber, J. Sabater-Mir, and M. Singh.

Demolombe, R. 2011. Transitivity and propagation of trust in information sources. An analysis in modal logic. In *Computational logic in multi-agent systems*, LNAI 6814, ed. J. Leite, P. Torroni, T. Agotnes, and L. van der Torre. Berlin/New York: Springer.

Demolombe, R., and C.-J. Liau. 2001. A logic of graded trust and belief fusion. In *Proceedings of 4th workshop on deception, fraud and trust*, Montreal, ed. C. Castelfranci and R. Falcone.

Jones, A.J., and M. Sergot. 1996. A formal characterisation of institutionalised power. *Journal of the Interest Group in Pure and Applied Logics* 4(3): 427–444

Jones, A.J.I. 2002. On the concept of trust. *Decision Support Systems* 33, 225–232.

Jones, A.J.I., and B.S. Firozabadi. 2001. On the characterisation of a trusting agent. Aspects of a formal approach. In *Trust and deception in virtual societies*, ed. C. Castelfranchi and Y.-H. Tan. Dordrecht/Boston: Kluwer Academic.

Lorini, E., and R. Demolombe. 2008. Trust and norms in the context of computer security: A logical formalization. In *Deontic logic in computer science*, LNAI 5076, ed. R. van der Meyden and L. van der Torre. Berlin/New York: Springer.

Pörn, I. 1977. Action theory and social science. Some formal models. *Synthese Library* 120.