

Reasoning About Trust: a Formal Logical Framework

Robert Demolombe

ONERA Toulouse
Toulouse, France

`Robert.Demolombe@cert.fr`

Abstract. There is no consensus about the definition of the concept of trust. In this paper formal definitions of different kinds of trust are given in the framework of modal logic. This framework also allows to define a logic for deriving consequences from a set of assumptions about trust.

Trust is defined as a mental attitude of an agent with respect to some property held by another agent. These properties are systematically analysed and we propose 6 epistemic properties, 4 deontic properties and 1 dynamic property.

In the second part of the paper more flexible notions of trust are introduced: qualitative graded trust, trust defined in terms of topics and conditional trust.

1 Introduction

The concept of trust is quite complex, it can be interpreted in many different ways and there is no consensus about its definition [10, 12, 14, 13, 15, 7, 11, 6]. That is the reason why we believe it is interesting to propose clear definitions that can be accepted, or rejected, but whose meaning is not a matter of discussion. That is the main purpose of this paper, and that is why the formal definitions are presented in formal logic.

In fact, we see three different kinds of problems related to trust. The first one is to define the facts that support trust. Trust can be supported by series of observations, or by reputation, or by an analysis of the situation. For example, one may trust a supplier about his capacity to deliver goods in time because he observed that in the past goods have been delivered in time, or because many people say that there are delivered in time, or because he knows the details of the supplier's organisation and he can conclude that there are good reasons to believe that they are delivered in time.

The second problem is to find the appropriate rules to derive consequences of a set of assumptions about trust, where the assumptions may be supported by the kinds of techniques we have mentioned just before.

The third problem is to use information about trust to take decisions. For example, if a manager has to assign a task to an employee he may select the employees in function of his trust about their capacities. We can say that in

general trust is used to complete our lack of information. It is better to have uncertain information derived from trust than to have nothing.

In this paper, in addition to a formal analysis of the definition of trust, we concentrate on reasoning about trust. That is we ignore the first and the third problems. Notice that the three problems can be investigated independently.

The starting point of our analysis is that trust is a mental attitude of an agent with respect to another agent¹. Here we take agent in a very broad sense. An agent may be a human agent or an artificial agent like a robot, or a sensor, or a program.

This attitude is a sort of belief about some property of an agent. Here we distinguish what we call belief, strong belief and knowledge. We say that an agent believes that some proposition holds if he has some justification for this belief and he believes that his justification may be wrong. For example, I believe that John is not at the University because his car is not at the parking. I know that sometimes John comes using a bicycle, therefore I may be wrong, but this justification is better than complete ignorance and it is enough to support my belief.

We say that an agent strongly believes that some proposition holds if he has some justification for his belief and he believes that this justification is true, though in reality it may be false. For example, I strongly believe that John is at the University because his car is at the parking and I believe he is the only driver of his car.

Finally, we say that an agent knows that some proposition holds if he has some justification for his knowledge and this justification is a true justification. For example, I know that John is at the University because I see him.

In the context of trust no information can be taken as a true information in the sense of knowledge. Then, if we say that some agent a trusts b with respect to some property that means that a strongly believes that b satisfies this property.

Now we can make more specific which kind of property may be trusted. In a previous work [4] we have only considered epistemic properties, like sincerity and credibility (see also [3]). Here we have also considered deontic properties like honesty, and dynamic properties like capacity.

In the section 2 of this paper we analyse epistemic properties, and in the section 3 we analyse deontic and dynamic properties. Finally, in the section 4 more flexible definitions are proposed. They include qualitative graded trust, synthetic trust about topics and conditional trust.

2 Trust about epistemic properties

To define epistemic properties we consider to what extent one of the following facts implies another one. These facts are denoted by:

¹ Here we take inspiration from a talk given by Andrew J.I. Jones in an informal workshop organised at the University of Lisbon in September 1997. One of his intuitive ideas is that “agent a trusts agent b if a believes that b will act in accordance with a norm which a believes b accepts”. More details can be found in [9].

p : it is the case that p .

$B_a p$: the agent a believes that p is the case.

$I_{a,b} p$ the agent a has informed the agent b that p is the case.

Then, the epistemic properties are defined as follows.

Sincerity. The agent b is sincere with regard to a for p iff if b informs a about p then b believes p . This property is formally represented by:

$$I_{b,a} p \rightarrow B_b p$$

Cooperativity. The agent b is cooperative with regard to a for p iff if b believes p then b informs a about p . That intuitively means that b does not hide p to a . This property is formally represented by:

$$B_b p \rightarrow I_{b,a} p$$

Credibility. The agent b is credible (or competent) about p iff if b believes p then p is the case. This property is formally represented by:

$$B_b p \rightarrow p$$

Vigilance. The agent b is vigilant about p iff if p is the case then b believes p . For example, in an airport the fact that an air traffic controller is vigilant about the fact that some aircraft has landed means that if the aircraft has landed the controller believes that it has landed. This property is formally represented by:

$$p \rightarrow B_b p$$

Validity. The agent b is valid with regard to a for p iff if b informs a about p then p is the case. For example, if the agent b is a sensor used to detect that a door is open, b is valid about the fact that the door is open means that if b sends the information that the door is open, then it is the case that the door is open. This property is formally represented by:

$$I_{b,a} p \rightarrow p$$

Completeness. The agent b is complete with regard to a for p iff if p is the case then b informs a about p . In the example of the sensor that means that if the door is open then the sensor sends the information that the door is open. This property is formally represented by:

$$p \rightarrow I_{b,a} p$$

Notice that all these definitions are defined in terms of implications. It may be tempting to define them in terms of conjunctions. For example, one could consider a definition of credibility of the form $(B_b p) \wedge p$. However, it is clear that the fact that b is **not** credible means that b believes p and p is not the case, which is represented by $(B_b p) \wedge \neg p$ which is logically equivalent to the negation of $B_b p \rightarrow p$, and it is not equivalent to the negation of $(B_b p) \wedge p$. The same

argument can be used to convince ourselves that the other definitions must also be defined with an implication.

It can also be noticed that all these properties are pairwise independent. For example, an agent may be sincere and not cooperative, and he may be as well cooperative and not sincere. However, three, or more, properties may be related. For example, validity is a consequence of sincerity and credibility, and completeness is a consequence of vigilance and cooperativity (see section 2.3).

2.1 Formal definitions of trust

Now we can define trust about these properties. We adopt the following notation:

K_ap : the agent a strongly believes that p is the case.

For the formal definitions we consider a modal propositional language with the modal operators we have presented above (see [2]).

We say, for example, that the agent a trusts b for his sincerity about p iff a strongly believes that b is sincere about p .

Trust is a mental attitude that is expressed in terms of strong beliefs because if the agent a does not believe that his justifications are true justifications, then that means that a may have some doubts about b 's sincerity, and in that case we cannot say that a really trusts b .

Trust of the agent a with regard to b about p for the property $prop$ is denoted by $Tprop_{a,b}p$. Then, trust for sincerity, cooperativity, credibility, vigilance, validity and completeness are respectively denoted by: $Tsinc_{a,b}(p)$, $Tcoop_{a,b}(p)$, $Tcred_{a,b}(p)$, $Tvigi_{a,b}(p)$, $Tval_{a,b}(p)$ and $Tcomp_{a,b}(p)$. Their formal definitions are:

$$Tsinc_{a,b}(p) \stackrel{\text{def}}{=} K_a(I_{b,a}p \rightarrow B_b p)$$

$$Tcoop_{a,b}(p) \stackrel{\text{def}}{=} K_a(B_b p \rightarrow I_{b,a} p)$$

$$Tcred_{a,b}(p) \stackrel{\text{def}}{=} K_a(B_b p \rightarrow p)$$

$$Tvigi_{a,b}(p) \stackrel{\text{def}}{=} K_a(p \rightarrow B_b p)$$

$$Tval_{a,b}(p) \stackrel{\text{def}}{=} K_a(I_{b,a} p \rightarrow p)$$

$$Tcomp_{a,b}(p) \stackrel{\text{def}}{=} K_a(p \rightarrow I_{b,a} p)$$

The formal definitions of the epistemic properties can be criticised because they are based on the material implication. That leads to a well known paradox. For example, for the definition of sincerity we can infer that b is sincere with regard to a for p in the situation where b has not informed a about p (i.e. when we have $\neg I_{b,a} p$). In other terms, an agent who says nothing is sincere.

The intuitive definition of sincerity should be that in every circumstances where b informs a about p then b believes p . This raises the well known issue of the formal representation of entailment which has no satisfactory solution.

To avoid too complex definitions we have accepted material implication, but since in the definition of trust the material implication is in the scope of the modal operator K_a the consequences are less dramatic. Nevertheless, if a strongly believes that it is not the case that b has informed a about p (i.e. $K_a(\neg I_{b,a}p)$) then we can infer that a trusts b about b 's sincerity. In the section 4, for the formalisation of graded trust we have used a conditional connective that avoids these problems.

2.2 Axiomatics

In addition to the axiom schemas and inference rules of propositional logic we have the following axiom schemas and inference rules.

The modal operator B_a obeys the system (KD) (see [2]). Then, in addition to the necessitation rule we have:

$$(K1) \quad B_a(p \rightarrow q) \rightarrow (B_ap \rightarrow B_aq)$$

$$(D1) \quad \neg(B_ap \wedge B_a\neg p)$$

The modal operator K_a obeys the system (KD) plus the axiom schema (KT).

$$(K2) \quad K_a(p \rightarrow q) \rightarrow (K_ap \rightarrow K_aq)$$

$$(D2) \quad \neg(K_ap \wedge K_a\neg p)$$

$$(KT) \quad K_a(K_ap \rightarrow p)$$

The schema (KT) intuitively means that a believes that what he believes is true. That is, a has no doubt about the truth of p . This schema characterises the notion of strong belief.

The modal operator $I_{a,b}p$ is not a normal operator it is a classical operator according to Chellas's classification [2]. For this operator we only have the rule of substitutivity of equivalent formulas.

$$(RE1) \quad \frac{\vdash p \leftrightarrow q}{\vdash I_{a,b}p \leftrightarrow I_{a,b}q}$$

These modal operators are not independent. In particular strong beliefs are a special kind of beliefs. Then, we have:

$$(KB) \quad K_ap \rightarrow B_ap$$

Also it is assumed that communication between the agents works perfectly well in the sense that each agent knows which message has been sent or has not been sent. Then we have:

$$(OBS1) \quad I_{b,a}p \rightarrow K_a(I_{b,a}p)$$

$$(OBS2) \quad \neg I_{b,a}p \rightarrow K_a(\neg I_{b,a}p)$$

2.3 Logical properties

As mentioned before trust about validity and about completeness are not independent of trust about other properties. We have:

$$Tsync_{a,b}(p) \wedge Tcred_{a,b}(p) \rightarrow Tval_{a,b}(p)$$

$$Tvig_{a,b}(p) \wedge Tcoop_{a,b}(p) \rightarrow Tcomp_{a,b}(p)$$

It is also interesting to see what can be inferred from the performance of the communication action $I_{b,a}p$ depending on what the agent a trusts. We have:

$$Tsync_{a,b}(p) \rightarrow (I_{b,a}p \rightarrow K_a(B_b p))$$

$$Tval_{a,b}(p) \rightarrow (I_{b,a}p \rightarrow K_a p)$$

$$Tcoop_{a,b}(p) \rightarrow (\neg I_{b,a}p \rightarrow K_a(\neg B_b p))$$

$$Tcomp_{a,b}(p) \rightarrow (\neg I_{b,a}p \rightarrow K_a(\neg p))$$

These properties show that we can infer information from the fact that b has not informed a about p in a similar way as we can infer information from the fact that b has informed a about p .

If we analyse the properties that relates trust for compound formulas of the form $p \wedge q$, $p \vee q$ or $\neg p$, in function of trust for p and q we see that these properties are very weak.

For the conjunction we have the following property if $prop = cred$ or $prop = vig$ ²:

$$Tprop_{a,b}(p) \wedge Tprop_{a,b}(q) \rightarrow Tprop_{a,b}(p \wedge q)$$

If for the modality $I_{b,a}$ we accept the axiom schema of monotonicity (M) $I_{b,a}(p \wedge q) \rightarrow I_{b,a}(p) \wedge I_{b,a}(q)$ the property holds for $prop = val$ and $prop = sinc$, and if we accept for this modality the axiom schema of closure (C) $I_{b,a}(p) \wedge I_{b,a}(q) \rightarrow I_{b,a}(p \wedge q)$ the property holds for $prop = comp$ and $prop = coop$. Acceptation of (M) and (C) comes to accept that saying p and q independently has the same consequences as saying $p \wedge q$.

Notice that the following implication never holds³.

$$Tprop_{a,b}(p \wedge q) \rightarrow Tprop_{a,b}(p) \wedge Tprop_{a,b}(q)$$

² We omit the proofs since they are very simple exercises.

³ This is a bit surprising. For example, if a trusts b for his credibility for $p \wedge q$ (i.e. $K_a(B_b(p \wedge q) \rightarrow (p \wedge q))$) we might expect that a trusts b for his credibility for p (i.e. $K_a(B_b p \rightarrow p)$). The reason why this property does not hold is that even if the set of worlds where we have $B_b(p \wedge q)$ is included in the set of worlds where we have $p \wedge q$, we cannot infer that the set of worlds where we have $B_b p$ is included in the set of worlds where we have p .

For the disjunction we have no property. That is $Tprop_{a,b}(p \vee q)$ does not imply $Tprop_{a,b}(p)$, and $Tprop_{a,b}(p)$ does not imply $Tprop_{a,b}(p \vee q)$.

For the negation the only property that holds is:

$$Tvgi_{a,b}(p) \rightarrow Tcred_{a,b}(\neg p)$$

If we accept the axiom schema $\neg B_a(p) \rightarrow B_a(\neg p)$ the property $Tcred_{a,b}(p) \rightarrow Tvgi_{a,b}(\neg p)$ holds. However, this axiom schema can be accepted only for the agents who have complete beliefs. If they have incomplete beliefs it leads to contradictions ⁴.

It is tempting to think that trust should be transitive. In fact, and this shows the benefit of formal definitions for reasoning about trust, this is not the case.

For example, if we assume that a trusts b about b 's sincerity with regard to a for p , and b trusts c about c 's sincerity with regard to b for p , should we infer that a trusts c for c 's sincerity with regard to a for p ? The answer is "no".

Indeed, in formal terms the question is: does $K_a(I_{b,a}p \rightarrow B_b p)$ and $K_b(I_{c,b}p \rightarrow B_c p)$ implies $K_a(I_{c,a}p \rightarrow B_c p)$? It is easy to define a counter example. Intuitively we can understand that from $K_a I_{c,a}p$ we can infer nothing using a 's trust about b and b 's trust about c .

However, if b says to a that c is sincere with regard to him (a) for p , and if a trusts that b is valid for what b said, then it can be inferred that a trusts that c is sincere with regard to a for p .

The formal proof below shows that from the hypothesis (H1) and (H2) we can infer $Tsinc_{a,c}p$.

$$(H1) Tval_{a,b}(I_{c,a}p \rightarrow B_c p)$$

$$(H2) I_{b,a}(I_{c,a}p \rightarrow B_c p)$$

$$(1) K_a(I_{b,a}(I_{c,a}p \rightarrow B_c p)) \rightarrow (I_{c,a}p \rightarrow B_c p) \text{ (H1), (Definitions)}$$

$$(2) I_{b,a}(I_{c,a}p \rightarrow B_c p) \text{ (H2)}$$

$$(3) I_{b,a}(I_{c,a}p \rightarrow B_c p) \rightarrow K_a I_{b,a}(I_{c,a}p \rightarrow B_c p) \text{ (OBS1)}$$

$$(4) K_a I_{b,a}(I_{c,a}p \rightarrow B_c p) \text{ (2), (3), (MP)}$$

$$(5) K_a(I_{c,a}p \rightarrow B_c p) \text{ (1), (4), (K)}$$

$$(6) Tsinc_{a,c}(p) \text{ (5), (Definitions)}$$

3 Trust about deontic and dynamic properties

To define deontic properties we analyse the possible links between the actions performed by the agents on one hand, and the obligations to perform actions on the other hand. The intuitive goal of this analysis is to define the deontic properties of the agents with regard to their fulfillment of a given regulation. We adopt the following notations:

$E_a p$: the agent a brings it about that p .

$O p$: it is obligatory that p .

As usual permission is defined in function of obligation. We have $P p \stackrel{\text{def}}{=} \neg O \neg p$, and $P p$ is read: it is permitted that p .

⁴ For example, if the agent a only believes $p \vee q$ we can infer $\neg B_a p$ and $\neg B_a q$, and then $B_a(\neg(p \vee q))$ which contradicts $B_a(p \vee q)$.

We first analyse the links between the fact that a brings it about that p and the fact that it is obligatory that a brings it about that p ⁵.

Obedience. The agent a is obedient for bringing it about that p iff if it is obligatory that a brings it about that p then a brings it about that p . In short terms, a does what it should do. This property is formally represented by:

$$OE_ap \rightarrow E_ap$$

Laziness. The agent a is lazy for p iff if a brings it about that p then it is obligatory that a brings it about that p . We call this property “laziness” because, by contraposition, if it is not obligatory that a brings it about that p then a does not bring it about that p . In other terms a only does what he is obliged to do. This property is formally represented by:

$$E_ap \rightarrow OE_ap$$

Notice that this definition is logically equivalent to $P\neg E_ap \rightarrow \neg E_ap$.

Now we analyse the links between the facts represented by E_ap and PE_ap .

Active. The agent a is active for p iff if it is permitted that a brings it about that p then a brings it about that p . Intuitively that means that a performs an action as soon as it is permitted to perform this action. This property is formally represented by:

$$PE_ap \rightarrow E_ap$$

Honesty. The agent a is honest for p iff if a brings it about that p then it is permitted that a brings it about that p . Then, an honest agent only does what it is permitted to do. This property is formally represented by:

$$E_ap \rightarrow PE_ap$$

Trust about obedience, laziness, active and honesty are respectively denoted by: $Tobed_{a,b}(p)$, $Tlazi_{a,b}(p)$, $Tacti_{a,b}(p)$ and $Thone_{a,b}(p)$. We have the formal definitions:

$$Tobed_{a,b}(p) \stackrel{\text{def}}{=} K_a(OE_ap \rightarrow E_ap)$$

$$Tlazi_{a,b}(p) \stackrel{\text{def}}{=} K_a(E_ap \rightarrow OE_ap)$$

$$Tacti_{a,b}(p) \stackrel{\text{def}}{=} K_a(PE_ap \rightarrow E_ap)$$

$$Thone_{a,b}(p) \stackrel{\text{def}}{=} K_a(E_ap \rightarrow PE_ap)$$

The axiomatics for reasoning about these properties is defined as follows.

⁵ Since we are interested by individual properties we have not considered the links between the facts represented by p and Op .

For the operator O we have adopted the simplest logic which is the standard deontic logic (more sophisticated deontic logics can be found in [1]). This logic is formalised by a (KD) system. Then, we have:

$$(K3) \quad O(p \rightarrow q) \rightarrow (Op \rightarrow Oq)$$

$$(D3) \quad \neg(Op \wedge O\neg p)$$

The E_a operator is a classical operator. It is assumed that this operator is a success operator. Then, we have the inference rule and axiom schema:

$$(RE2) \quad \frac{\vdash p \leftrightarrow q}{\vdash E_ap \leftrightarrow E_aq}$$

$$(T) \quad E_ap \rightarrow p$$

To analyse dynamic properties we have introduced the modal operator H_a (see [8]). The sentence H_ap can be read: the agent a attempts to bring it about that p . We consider below the links between the facts represented by p and H_ap .

Ability. The agent is able to bring it about that p iff if a is in the context c that makes possible to bring it about that p ⁶ and a attempts to bring it about that p , then p is obtained. This property is formally represented by:

$$c \rightarrow (H_ap \rightarrow p)$$

Of course, in this definition p denotes a propositional constant, not a propositional variable like in the axiom schema (T).

The property formally represented by $p \rightarrow H_ap$ means that if p is the case then the agent a attempts to bring it about that p . This corresponds to no property which has an intuitive meaning.

Trust about ability is denoted by $Tabil_{a,b}(p, c)$ and it is defined by:

$$Tabil_{a,b}(p, c) \stackrel{\text{def}}{=} K_a(c \rightarrow (H_ap \rightarrow p))$$

The modal operator H_a is a classical operator that obeys the inference rule:

$$(RE3) \quad \frac{\vdash p \leftrightarrow q}{\vdash H_ap \leftrightarrow H_aq}$$

4 More flexible definitions of trust

4.1 Graded trust

In section 2 the definitions of trust have the general form: $K_a(\phi_b \rightarrow \psi_b)$. For example, for sincerity we have $K_a(I_{b,a}p \rightarrow B_bp)$. The intended meaning is that

⁶ We would like to thank the anonymous referee who pointed out the necessity for an agent to be in a given context to be able to exercise his ability.

a strongly believes that in all the circumstances where we have ϕ_b we also have ψ_b . That is, the set of worlds (compatible with what a strongly believes) where we have ϕ_b is included in the set of worlds where we have ψ_b . In most of the real situations our trust is less rigid and we strongly believe that the first set is “more or less” included in the second set.

To formalise this notion of “more or less” included we have introduced the connective \Rightarrow_i which is indexed by the “level of inclusion” i . The formula $\phi \Rightarrow_i \psi$ intuitively means that ϕ entails ψ at the level i .

The semantics of this connective is defined in a similar way as the conditional connective defined in [2]. For a given world w of a given model M we have:

$$M, W \models \phi \Rightarrow_i \psi \text{ iff } f_i(w, |\phi|_M) = |\psi|_M$$

where $|\phi|_M$ denotes the set of worlds w' such that $M, w' \models \phi$ ⁷.

In a model a function f_i has to be defined for each level i . A very particular case might be to define f_i like a probability, that is to have: $f_i(w, |\phi|_M) = |\psi|_M$ iff $\frac{card|\phi \wedge \psi|_M}{card|\phi|_M} = i$, where $card|\phi|_M$ denotes the cardinality of the set $|\phi|_M$. However, in many applications such quantitative levels have no intuitive meaning and we propose to have only qualitative levels for which is defined a partial order relation denoted by $i < j$.

For the different kinds of properties *prop* we denote the level of trust by $Tprop_{a,b}^i(p)$ and we have definitions of the form:

$$Tprop_{a,b}^i(p) \stackrel{\text{def}}{=} K_a(\phi_b \Rightarrow_i \psi_b)$$

For example, in the case of sincerity we have: $Tsinc_{a,b}^i(p) \stackrel{\text{def}}{=} K_a(I_{b,a}p \Rightarrow_i B_b p)$.

We think that a possible method to assign an intuitive meaning to each level is to assign some particular agent to each level and to use these agents as references for each level.

For example, if the agent r is used as a reference for the level i with regard to some proposition p , it is possible to assign the level i to another agent b if a trusts b in the same way (at the same level) as a trusts r . More formally, if r is the reference for the level i we should have $Tprop_{a,b}^i(p)$ iff $Tprop_{a,r}^i(p)$.

The consequences of graded trust are graded beliefs. We adopt the notation:

$B_a^i p$: the agent a believes at the level i that p is the case.

If we have $K_a(\phi \rightarrow \psi)$ from $K_a\phi$ we can infer $K_a\psi$, but if we have $K_a(\phi \Rightarrow_i \psi)$ from $K_a\phi$ we can only infer the weaker consequence $B_a^i\psi$. Then, we have the following axiom schema:

⁷ In the definition of conditionals in [2] we have in the satisfiability condition $f_i(w, |\phi|_M) \subseteq |\psi|_M$. A consequence of this definition is that we have $\models \psi \rightarrow \psi' \Rightarrow \models (\phi \Rightarrow_i \psi) \rightarrow (\phi \Rightarrow_i \psi')$ (see the rule RCK in [2]) and we do not want this property for the connective \Rightarrow_i . Indeed, in $\phi \Rightarrow_i \psi$ the connective \Rightarrow_i expresses to what extent the fact that a world is in $|\phi|_M$ entails the fact that this world is in $|\psi|_M$. If we accept the definition given in [2] from $\phi \Rightarrow_i \psi$ we can infer $\phi \Rightarrow_i \psi \vee \phi$, and in general the “strength” of the entailment in $\phi \Rightarrow_i \psi$ is not the same as its strength in $\phi \Rightarrow_i \psi \vee \phi$.

$$(K_i) \quad K_a(\phi \Rightarrow_i \psi) \rightarrow (K_a\phi \rightarrow B_a^i\psi)$$

The links between graded beliefs and strong beliefs are defined by the following schemas:

$$\begin{aligned} &\text{If } j \leq i, B_a^i\phi \rightarrow B_a^j\phi. \\ &\text{For every } i, K_a\phi \rightarrow B_a^i\phi. \end{aligned}$$

4.2 Trust with respect to topics

In general trust is not specific to a proposition p . Rather an agent trusts another agent for all the propositions related to a given topic. For instance, the agent a may trust b with respect to his validity for all the propositions that are about the topic “nuclear”.

In [5] Demolombe and Jones have defined the property:

$A(t, “p”)$: the sentence named by “p” is about the topic t , and they have defined a logic for reasoning about sentences of the form $A(t, “p”)$. This logic is weaker than a classical logic. The only inference rule is:

$$\text{if } Var(p) = Var(q), \quad \frac{\vdash p \leftrightarrow q}{\vdash A(t, “p”)\leftrightarrow A(t, “q”)}$$

where $Var(p)$ denotes the set of propositional variables in p .

We denote by $Tprop_{a,b}(t)$ the fact that a trusts b with respect to $prop$ for the topic t and we have the formal definition:

$$Tprop_{a,b}(t) \stackrel{\text{def}}{=} \forall “p” (A(t, “p”) \rightarrow Tprop_{a,b}(p))$$

Moreover, it is quite natural to define a structure over the set of topics with the notion of specificity. We denote by $t' \text{ isa } t$ the fact that the topic t' is more specific than t . For example, the topic “nuclear weapon” is more specific than the topic “nuclear”. According to this structure we adopt the axiom schema:

$$t' \text{ isa } t \rightarrow (A(t', “p”) \rightarrow A(t, “p”))$$

4.3 Conditional trust

There are many situations where an agent trusts another agent only in some particular circumstances. For instance, let us suppose that the agent b is the sensor that detects that the door is open. An agent a may trust b with regard to his completeness for the fact that the door is open (represented by p) only if the electric power is on (represented by q).

This kind of trust is formally represented by $K_a(q \rightarrow (p \rightarrow I_{a,b}p))$.

In general, the fact that a trusts b for p in the circumstances represented by q is denoted by $Tprop_{a,b}(p|q)$ and we have the formal definition:

$$Tprop_{a,b}(p|q) \stackrel{\text{def}}{=} K_a(q \rightarrow prop(p))$$

where $prop(p)$ is any property about p of the kind that we have seen in the section 2 or 3.

It is worth noting that conditional trust would not be correctly represented by: $q \rightarrow K_a(prop(p))$, because what the agent a trusts depends on what he strongly believes (i.e. $K_a q$) and not on the fact that really holds (i.e. q).

We have the intuitive property:

$$Tprop_{a,b}(p|q) \rightarrow (K_a q \rightarrow Tprop_{a,b}(p))$$

5 Conclusion

We have defined trust as a mental attitude of an agent with respect to another agent which has been called strong belief. We have shown that this property may be an epistemic, a deontic or a dynamic property.

These properties have been systematically analysed in the framework of modal logic. The technic was to consider facts represented by p , $B_a p$ and $I_{b,a} p$, for epistemic properties, facts represented by $E_a p$, $OE_a p$ and $PE_a p$, for deontic properties, and facts represented by $H_a p$ and p for dynamic properties. An axiomatic characterisation has been given for the modal operators involved in these definitions.

This analysis has allowed to “rediscover” some intuitive properties and to exhibit some of them that are non trivial.

At the beginning trust is defined for a specific fact p and we only consider two situations: an agent trusts, or does not trust, another agent. In the second part of the paper have been introduced more flexible definitions that are closer to the definitions used in practical applications. Graded trust allows to represent several qualitative levels of trust. Trust about topics allows to define more generic trust. Finally, conditional trust allows to relativize trust to some particular circumstances. These extensions of the notion of trust could be easily combined to define, for example, graded trust defined in terms of topics.

References

1. J. Carmo and A.J.I. Jones. Deontic Logic and Contrary-to Duties. In D. Gabbay, editor, *Handbook of Philosophical Logic (Rev. Edition)*. Reidel, to appear.
2. B. F. Chellas. *Modal Logic: An introduction*. Cambridge University Press, 1988.
3. R. Demolombe. Validity Queries and Completeness Queries. In *Proc. of 9th International Symposium on Methodologies for Intelligent Systems*, 1996.
4. R. Demolombe. To trust information sources: a proposal for a modal logical framework. In C. Castelfranchi and Y-H. Tan, editors, *Trust and Deception in Virtual Societies*. Kluwer Academic Publisher, 2001.
5. R. Demolombe and A.J.I. Jones. On sentences of the kind “sentence “p” is about topic “t””: some steps toward a formal-logical analysis. In H-J. Ohlbach and U. Reyle, editor, *Logic, Language and Reasoning. Essays in Honor of Dov Gabbay*. Kluwer Academic Press, 1999.

6. G. Elofson. Developing trust with intelligent agents: an explanatory study. In C. Castelfranchi and Y-H. Tan, editors, *Trust and Deception in Virtual Societies*. Kluwer Academic Publisher, 2001.
7. R. Falcone and C. Castelfranchi. Social trust: a cognitive approach. In C. Castelfranchi and Y-H. Tan, editors, *Trust and Deception in Virtual Societies*. Kluwer Academic Publisher, 2001.
8. A.J.I. Jones. A logical framework. In J. Pitt, editor, *The Open Agent Society*. John Wiley and Sons.
9. A.J.I. Jones. *Communication and meaning: An essay in applied modal logic*. Synthese Library. Reidel, 1983.
10. A.J.I. Jones. On the concept of trust. *Decision Support Systems*, 33, 2002.
11. A.J.I. Jones and B.S. Firozabadi. On the characterisation of a trusting agent. Aspects of a formal approach. In C. Castelfranchi and Y-H. Tan, editors, *Trust and Deception in Virtual Societies*. Kluwer Academic Publisher, 2001.
12. F.J. Lerch and M.J. Prietula. How do we trust machine advice. In *Third International Conference on Human-Computer Interaction*, 1989.
13. S. March. Trust in distributed Artificial Intelligence. *Lectures Notes in Computer Science*, 830, 1994.
14. P.Oerbaek. Can you Trust your data. *Lectures Notes in Computer Science*, 915, 1995.
15. K. Thompson. Reflections on Trusting Trust. *Communications of ACM*, 27(18), 1984.