Trust Dynamics: How Trust is influenced by direct experiences and by Trust itself[1]

Rino Falcone
ISTC -CNR
Roma, Italy
r.falcone@istc.cnr.it

Cristiano Castelfranchi
ISTC -CNR
Roma, Italy
c.castelfranchi@istc.cnr.it

Abstract

In this paper we will examine two main aspects of trust dynamics:

a) How direct experiences involving trust, with their successes or failures, influence the future trust of an agent about similar facts. We challenge the trivial idea that always success increases trust while failure decreases it. Of course, this primitive view cannot be avoided till Trust is modeled just as a simple index, a dimension, a number; for example reduced to mere subjective probability. We claim that a cognitive attribution process is needed in order to update trust on the basis of an 'interpretation' of the outcome of A's reliance on B and of B's performance (failure or success).

b) How the fact that A trusts B and relies on it in situation Ω can actually (objectively) influence B's trustworthiness in the Ω situation. Either trust is a self-fulfilling prophecy that modifies the subjective probability of the predicted event; or it is a self-defeating strategy by negatively influencing the events.

These phenomena are very crucial in human societies (states, market, groups), but also in computer mediated Organizations, Interactions (EC), Cooperation (CSCW) and even in Multi-Agent Systems with autonomous agents. We present a formal model of these dynamic non-trivial aspects.

1. Introduction

Trust is becoming one of the main subjects of study in the Information Society Technologies. In fact, the success of computer supported society in which humans have always more to cope with unusual entities: new kind of environments, of procedures, of interactions and partners (smart physical environment, virtual reality, virtual organization, artificial agents, and so on), is mainly dependent from the trust and confidence in this new kind of environments and societies.

We have to better understand trust both as a theoretical concept and as a useful tool for simplifying the world and for coping with risks in it and uncertainty: trust in the computational infrastructure; trust in potential partners, trust in information sources, data, mediating agents, trust in personal assistants; trust in other agents and processes. Security measures are not enough; interactivity and knowledge ability are not enough; the problem is how to build in users and agents trust and how to maintain it.

Trust is a dynamic phenomenon in its intrinsic nature. Trust changes with experience, with the modification of the different sources it is based on, with the emotional state of the trustier, with the modification of the environment in which the trustee is supposed to perform, and so on. In other words, being trust an attitude depending from dynamic phenomena, as a consequence it is itself a dynamic entity.

There are many studies in literature dealing with the dynamics of trust [1,2,3,4]. We are interested to analyze two main basic aspects of this phenomenon:

i) The traditional problem of the trust reinforcement on the basis of the successful experiences (and vice versa, its decreasing in case of failures);

ii) The fact that in the same situation *trust is influenced by trust* in several rather complex ways.

The first case (i) considers the well known phenomenon about the fact that trust evolves in time and has a history, that is A's trust in B depends on A's previous experience and learning with B itself or with other (similar) entities. In §3 we will analyze this case and will also consider some not so easily predictable results in which trust in the trustee decreases with positive experiences (when the trustee realizes the delegated task) and increases with negative experiences (when the trustee does not realize the delegated task).

Since trust is not simply an external observer's prediction or expectation about a matter of fact, we will also consider in the second case (ii) the fact that in one and the same situation *trust is influenced by trust* in several rather complex ways. We have analyzed in

[1] This paper has been partially funded by the TICCA Project: joint research venture between the Italian National Research Council and Provincia Autonoma di Trento; and by the Progetto Miur Cofin 2003 "Fiducia e diritto nella società dell'informazione".

AAMAS'04, July 19-23, 2004, New York, New York, USA.
Copyright 2004 ACM 1-58113-864-4/04/0007...$5.00

another paper part of this phenomenon [4]: *How trust creates a reciprocal trust, and distrust elicits distrust*; but also vice versa: how *A*'s trust in *B* could induce lack of trust or distrust in *B* towards *A*, while *A*'s diffidence can make B more trustful in *A*.

In this paper we will examine in the §4 an interesting aspect of trust dynamics: *How the fact that A trusts B and relies on it in situation Ω can actually (objectively) to influence B's trustworthiness in the Ω situation*. Either trust is a self-fulfilling prophecy that modifies the probability of the predicted event; or it is a self-defeating strategy by negatively influencing the events. And also how *A* can be aware of (and takes into account) the effect of its own decision in the very moment of that decision.

As we have argued in previous works [4, 5, 6] and we will resume in §2, trust and reliance/delegation are strictly connected phenomena: trust could be considered as the set of mental components a delegation action is based on. In the analysis of trust dynamic, we have also to consider the role of delegation (*weak*, *mild* and *strong* delegation) [7].

2. Socio-Cognitive Model of Trust

The Socio-Cognitive model of trust [5] is based on a portrait of the mental state of trust in cognitive terms (beliefs, goals). This is not a complete account of the psychological dimensions of trust: it represents the most explicit (reason-based) and conscious form. The model does not account for the more implicit forms of trust (for example trust by default, not based upon explicit evaluations, beliefs, derived from previous experiences or other sources) or for the affective dimensions of trust, based not on explicit evaluations but on emotional responses and an intuitive, unconscious appraisal.

The word *trust* means different things, but they are systematically related with each other. In particular, three crucial concepts have been recognized and distinguished not only in natural language but also in the scientific literature. Trust is at the same time:

- A mere *mental attitude* (prediction and evaluation) towards another agent, a simple *disposition*;

- A *decision* to rely upon the other, i.e. an *intention* to delegate and trust, which makes the trustier "vulnerable";

- A *behaviour*, i.e. the intentional *act* of trusting, and the consequent *relation* between the trustier and the trustee.

In each of the above concepts, different sets of cognitive ingredients are involved in the trustier's mind. The model is based on the BDI (Belief-desire-intention) approach for mind modelling, that is inspired by Bratman's philosophical model [8]. First of all, in the trust model only an agent endowed with both goals and beliefs can *trust* another agent. Let us consider the trust of the agent *X* towards another agent *Y* about the (*Y*'s) behaviour/action □ relevant for the result (goal) *g* when:

X is the (relying) agent, who feels trust; it is a cognitive agent endowed with internal explicit goals and beliefs (*the trustier*); *Y* is the agent (or entity), which is trusted (*the trustee*). *X* trusts *Y* about *g*/□ and for *g*/□.

For all the three notions of trust above defined (trust disposition, decision to trust, and trusting behaviour) we claim that someone trusts someone other only relatively to some goal (here goal is intended as the general, basic teleonomic notion, any motivational representation in the agent: desires, motives, will, needs, objectives, duties, utopias, are kinds of goals). An unconcerned agent does not really trust: he just has opinions and forecasts. Second, trust itself consists of beliefs, in particular: evaluations and expectations.

Since *Y*'s action is useful to *X* (trust disposition), and *X* has decided to rely and depend on it (decision to trust), this means that *X* might delegate (act of trusting) some action/goal in his own plan to *Y*. This is the strict relation between trust disposition, decision to trust, and delegation.

The model includes two main basic beliefs (we are considering the trustee as a cognitive agent too):

- *Competence Belief*: a *sufficient evaluation* of *Y*'s abilities is necessary, *X* should believe that *Y* is useful for this goal of its, that *Y* can produce/provide the expected result, that *Y* can play such a role in *X*'s plan/action.

- *Willingness Belief*: *X* should think that *Y* not only is able and can do that action/task, but *Y* actually will do what *X* needs (under given circumstances). This belief makes the trustee's behaviour predictable.

Another important basic belief for trust is:

- *Dependence Belief*: *X* believes -to trust *Y* and delegate to it- that either *X* needs it, *X* depends on it (*strong dependence*), or at least that it is better to *X* to rely rather than do not rely on it (*weak dependence*). In other terms, when *X* trusts someone, *X* is in a strategic situation: *X* believes that there is interference and that his rewards, the results of his projects, depend on the actions of another agent *Y*.

Obviously, the willingness belief hides a set of other beliefs on the trustee's reasons and motives for helping. In particular, *X* believes that *Y* has some motives for helping it (for adopting its goal), and that these motives will probably prevail -in case of conflict- on other motives. Notice that motives inducing adoption are of several different kinds: from friendship to altruism, from morality to fear of sanctions, from exchange to common goal (cooperation), and so on.

From the point of view of the dynamic studies of trust, it is relevant to underline how the above basic beliefs might change during the same interaction or during several interactions: for example could change the abilities of the trustee or his/her reasons/motives for willing (and/or the trustier's beliefs on them); or again it might change the dependence relationships between the trustier and the trustee.

AAMAS'04, July 19-23, 2004, New York, New York, USA.

Another important characteristic of the socio-cognitive model of trust is the distinction between trust ‘in’ someone or something that has to act and produce a given performance thanks to its *internal characteristics*, and the global trust in the global event or process and its result which is also affected by *external factors* like opportunities and interferences.

Trust in *Y* (for example, ‘social trust’ in strict sense) seems to consists in the two first prototypical beliefs/evaluations identified as the basis for reliance: *ability/competence* (that with cognitive agents includes knowledge and self-confidence*)*, and *disposition* (that with cognitive agents is based on willingness, persistence, engagement, etc.). Evaluation about external opportunities is not really an evaluation about *Y* (at most the belief about its ability to recognize, exploit and create opportunities is part of our trust ‘in’ *Y*). We should also add an evaluation about the probability and consistence of obstacles, adversities, and interferences.

Trust can be said to consist of, or better to (either implicitly or explicitly) imply the *subjective probability* of the successful performance of a given behaviour □, and it is on the basis of this subjective perception/evaluation of risk and opportunity that the agent decides to rely or not *Y*. However, the probability index is based on, derives from those beliefs and evaluations. In other terms the global, final subjective probability of the realization of the goal *g*, i.e. of the successful performance of □, should be decomposed into the expectation of *Y* performing the action well (*internal attribution*) and the expectation of having the appropriate conditions (*external attribution*) for the performance and for its success, and of not having interferences and adversities (*external attribution*). This decomposition is important because:

The trustier’s decision might be different with the same global probability or perceived risk, depending on its composition (for example for personality factors);

Trust composition (internal Vs external) produces completely different intervention strategies: to manipulate the external variables (circumstances, infrastructures) is completely different than manipulating internal parameters.

Let us now introduce a few formal constructs. We define Act={$□_1$,..,$□_n$} be a finite set of *actions*, and Agt={X, Y, A, B,..} a finite set of *agents*. Each agent has an action repertoire, a plan library, resources, goals (in general by "goal" we mean a partial situation (a subset) of the "world state"), beliefs, motives, etc.[2]

We consider the action/goal pair *□=(□,g)* as the real object of delegation, and we will call it ‘task’. Then by means of □, we will refer to the action (□), to its resulting world state (*g*), or to both.

Given an agent *X* and a situational context Ω (a set of propositions describing a state of the world), we define as trustworthiness of *X* about □ in Ω (called *trustworthiness (X □Ω)*), the objective probability that *X* will successfully execute the task □ in context Ω. This objective probability is in terms of our model computed on the basis of some more elementary components:

- A *degree of ability* (*DoA*, ranging between O and 1, indicating the level of *X*'s ability about the task □); we can say that it could be measured as the number of *X*'s successes (*s*) on the number of *X*'s attempts (*a*): *s/a*, when *a* goes to ∞; and

- A *degree of willingness* (*DoW*, ranging between O and 1, indicating the level of *X*'s intentionality/persistence about the task □); we can say that it could be measured as the number of *X*'s (successfully or unsuccessfully) performances (*p*) of that given task on the number of times *X* declares to have the intention (*i*) to perform that task: *p/i*, when *i* goes to ∞; we are considering that an agent declares its intention each time it has got one. So, in this model we have that:

$$Trustworthiness\ (B\ □\ \Omega) = F(\ DoA_{B\ □\ \Omega}\ ,\ \ DoW_{B\ □\ \Omega})$$

Where *F* is in general a function that preserves monotonicity, and ranges in (0,1): for the purpose of this work it is not relevant to analyze the various possible models of the function *F*. We have considered this probability as *objective* (absolute, not from the perspective of another agent) because we hypothesize that it measures the real value of the *X*'s trustworthiness; for example, if *trustworthiness(X □Ω)*=0.80, we suppose that in a context Ω, 80% of times *X* tries and succeed in executing □

As the reader can see we have not considered the opportunity dimension: the external conditions allowing or inhibiting the realization of the task.

3. Experience as a Reasoning Process: Causal Attribution for Trust

It is commonly accepted [1,2,3,9] and discussed in another our work [10] that one of the main sources of trust is the direct experience. Generally, in this kind of experiences to each success of the trustee corresponds an increment in the amount of the trustier’s trust towards it, and vice versa, to every trustee’s failure corresponds a reduction of the trustier’s trust towards the trustee itself. There are several ways in which this qualitative model could be implemented in a representative dynamic function (linearity or not of the function; presence of possible thresholds (under a minimum threshold there is no trust, or vice versa, over a maximum threshold there is full trust), and so on).

This view is very naïve, neither very explicative for humans and organizations, nor useful for artificial

[2] We assume that *to delegate an action necessarily implies delegating some result of that action.* Conversely, *to delegate a goal state always implies the delegation of at least one action (possibly unknown to Y) that produces such a goal state as result.*

systems, since it is unable to adaptively discriminate cases and reasons of failure and success. However, this primitive view cannot be avoided till Trust is modeled just as a simple index, a dimension, a number; for example, reduced to mere subjective probability. We claim that a cognitive attribution process is needed in order to update trust on the basis of an '*interpretation*' of the outcome of *A*'s reliance on *B* and of *B*'s performance (failure or success). In doing this a cognitive model of Trust – as we have presented – is crucial. In particular we claim that the effect of both *B*'s failure or success on *A*'s Trust in *B* depends on *A*'s *'causal attribution'* [11] of the event. Following 'causal attribution theory' any success or failure can be either ascribed to factors *internal* to the subject, or to environmental, *external* causes, and either to *occasional* facts, or to *stable* properties (of the individual or of the environment). So, there are four possible combinations: *internal* and *occasional*; *internal* and *stable*; *external* and *occasional*; *external* and *stable*.

The cognitive, emotional, and practical consequences of a failure (or success) strictly depends on this causal interpretation. For example –psychologically speaking– a failure will impact on the self-esteem of a subject only when attributed to *internal* and *stable* characteristics of the subject itself. Analogously, a failure is not enough for producing a crisis of trust; it depends on the causal *interpretation* of that outcome, on its attribution (the same for a success producing a confirmation or improvement of trust). In fact, we can say that a first qualitative result of the causal interpretation can be resumed in the following flow chart (Figure 1).

Since in agent-mediated human interaction (like CSCW or EC) and in cooperating autonomous MAS it is fundamental to have a theory of, and instruments for, **'Trust building**' we claim that a correct model of this process will be necessary and much more effective.

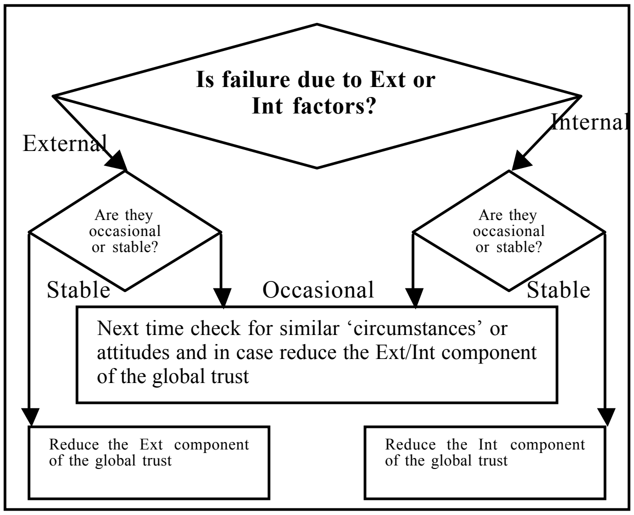
Is failure due to Ext or Int factors?
External
Internal
Are they occasional or stable?
Are they occasional or stable?
Stable
Occasional
Stable
Next time check for similar 'circumstances' or attitudes and in case reduce the Ext/Int component of the global trust
Reduce the Ext component of the global trust
Reduce the Int component of the global trust

Figure 1

Let's first present a basic model (which exploits our cognitive analysis of Trust attitude and 'causal attribution theory' which are rather convergent), and later discuss possible more complex dynamics. The following analysis takes into account the stable facts.

We consider a general function by which the agent *A* evaluates its own trust (*degree of trust*) in agent *B* about the task τ (to be performed) in the environment Ω ($DoT_{A,B,\tau,\Omega}$): $DoT_{A,B,\tau,\Omega} = f(DoA_{A,B,\tau}, DoW_{A,B,\tau}, e(\Omega))$)

Where: f (like F) is a general function that preserves monotonicity. In particular, $DoA_{A,B,\tau}$ is the *B*'s degree of ability (in *A*'s opinion) about the task τ, $DoW_{A,B,\tau}$ is the *B*'s degree of motivational disposition (in *A*'s opinion) about the task τ (both $DoA_{A,B,\tau}$ and $DoW_{A,B,\tau}$ are evaluated in the case in which *B* would try to achieve that task in a standard environment: an environment with the commonly expected and predictable features); *e(Ω)* takes into account the part of the task not directly performed by *B* (this part cannot be considered as a separated task but as integrating part of the task and without which the same task cannot be considered as complete) and the hampering or facilitating conditions of the specific environment. In a simplified analysis of these three sub-constituents *($DoA_{A,B,\tau}$ (Abilities), $DoW_{A,B,\tau}$ (Motivations),* and *e(Ω) (Environment))* of the *A*'s degree of trust, we have to consider the different possible dependencies among these factors:

i) we always consider *Abilities* and *Motivations* as independent to each other (for sake of simplicity);

ii) case in which *Abilities* and *Motivations* are both independent from the *Environment*: it is the case in which there is a part of the task performed (activated, supported etc.) from the *Environment* and, at the same time, both *Abilities* and *Motivations* cannot influence this part of the task. Consider for example, the task of urgently deliver an important apparatus to a scientific laboratory in another town. Suppose that this apparatus could be sent by using any service of delivery (public, private, fast or normal, and so on) so that a part of the task (to materially bring the apparatus) is independent (once made the choice) from the actions of the trustee.

iii) case in which *Abilities* and *Environment* are dependent with each other. We have two sub-cases: first, the *Environment* favours or disfavours the *B*'s *Abilities* (useful for the task achievement); second, the *B*'s *Abilities* can modify some of the conditions of the *Environment* (both these sub-cases could be known or not before the task assignment).

iv) case in which *Motivations* and *Environment* are dependent with each other. Like for the case (iii), there are two sub-cases: first, the *Environment* influences the *B*'s *Motivations* (useful for the task achievement); second, the *B*'s *Motivations* can modify some of the conditions of the *Environment* (both these sub-cases could be known or not before the task assignment).

Given this complex set of relationships among the various sub-constituents of trust, it is clear that analyzing

AAMAS'04, July 19-23, 2004, New York, New York, USA.
Copyright 2004 ACM 1-58113-864-4/04/0007...$5.00

and understanding the different role played by each ingredient in the specific performance (the specific experiential event), a trustier well informed and supplied with an analytic apparatus (a socio-cognitive agent), could evaluate which ingredients performed well and which failed.

Let us start from the case in which *Abilities* and *Motivations* both are considered as composed of internal properties and independent from the *Environment* (case (ii)). After an experiential event the trustier could verify:

1) Actual(*DoA, DoW*) – Expected(*DoA, DoW*) > 0

2) Actual(*DoA, DoW*) – Expected(*DoA, DoW*) < 0

3) Actual(*e(*Ω*)*) – Expected(*e(*Ω*)*) > 0

4) Actual(*e(*Ω*)*) – Expected(*e(*Ω*)*) < 0

In (1) and (2) both the trustee (int-trust) and the environment (ext-trust) are more trustworthy than expected; vice versa, in (2) and (4) they are both less trustworthy than expected.

In Table1 are shown all the possible combinations.

	Success attribution	Failure attribution
□(int-trust) > 0 □(ext-trust) > 0	**A** More Int-trust; More ext-trust	**A'** More Int-trust; More ext-trust
□(int-trust) > 0 □(ext-trust) < 0	**B** More Int-trust; Less ext-trust	**B'** More Int-trust; Less ext-trust
□(int-trust) < 0 □(ext-trust) > 0	**C** Less Int-trust; More ext-trust	**C'** Less Int-trust; More ext-trust
□(int-trust) < 0 □(ext-trust) < 0	**D** Less Int-trust; Less ext-trust	**D'** Less Int-trust; Less ext-trust

Table 1

Where: "More Int-trust" ("Less Int-trust") means that the trustier after the performance considers the trustee more (less) trustworthy than before it; "More Ext-trust" ("Less Ext-trust") means that the trustier after the performance considers the environment more (less) trustworthy than before it.

Particularly interesting are the cases:

(B) in which even if the environment is less trustworthy than expected, the better performance of the trustee produces a global success performance.

(C) in which even if the trustee is less trustworthy than expected, the better performance of the environment produces a global success performance. *An interesting case in which is possible decreasing the trust in the trustee even in presence of a success.*

(D and A') In which expectations do not correspond with the real trustworthiness necessary for the task (too high in D and too low in A'). These cases are not possible if the trustier has a correct perception of the necessary level of trustworthiness for that task (as we suppose in the other cases in Table1).

(B') in which even if the trustee is more trustworthy than expected (so increases the trust in it), the unexpected (at least for the trustier) difficulties introduced by the environment produces a global failure performance. *An interesting case in which is possible increasing the trust in the trustee even in presence of a failure.*

Again more complex is the case in which there is dependence between the internal properties and the environment (cases (iii) and (iv). In this case, in addition to the introduced factors □(int-trust) and □(ext-trust), we have to consider also the factor taking into account the possible influences between internal and external factors. We consider these influences as not expected from the trustier in the sense that the expected influences are integrated directly in the internal or external factors. Only as an example, we can consider the case of a violinist. We generally trust him to play a good performance; but suppose he has to do the concert in an open environment and the weather conditions are particularly bad (very cold): may be that these conditions can modify the specific hand abilities of the violinist and of his performance; at the same way, it is possible that a special conflicting audience could modify his willingness and consequently again his performance.

4. Changing the Trustee's Trustworthiness

In this section we are going to analyze how a delegation action (corresponding to a decision making based on trust in a specific situational context) could change the trustworthiness of the delegated agent (*delegee*).

4.1 The Case of Weak Delegation

We call *weak delegation* (and express this with *W-Delegates(A B □)*) the reliance simply based on exploitation for the achievement of the task. In it there is no agreement, no request or even (intended) influence: *A is just exploiting in its plan a fully autonomous action of B*. For a more complete discussion on the mental ingredients of the weak delegation see [7].

We hypothesize that in weak delegation (as in any delegation) there is a decision making based on trust and in particular there are two specific beliefs of *A*: - belief1: if *B* makes the action then *B* has a successful performance; - belief2: *B* intend to do the action.

As showed in §3 the trustworthiness of *B* by *A* is:

$DoT_{A,B,\square,\Omega} = f(DoA_{A,B,\square}, DoW_{A,B,\square}, e(\Omega))$

For sake of semplicity we assume that:

$DoT_{A,B,\square,\Omega} \equiv trustworthiness\ (B\ \square\ \Omega)$

in words: *A* has a perfect perception of the *B*'s trustworthiness. The interesting case in weak delegation is

AAMAS'04, July 19-23, 2004, New York, New York, USA.
Copyright 2004 ACM 1-58113-864-4/04/0007...$5.00

when: $Bel(A\ \neg Bel(B\ W\text{-}Delegates(A,B,\tau)))\ \wedge\ Bel(B\ W\text{-}Delegates(A,B,\tau))$[3]

in words, there is a weak delegation by *A* on *B* and *B* is aware of it (while *A* believes that *B* is not conscious).

The first belief is very often true in weak delegation, while the second one is necessary in the case we are going to consider. If *Bel (B W-Delegates (A B τ)),* this belief could change the *B*'s trustworthiness, either because *B* will adopt *A*'s goal and accepts such a exploitation, or because *B* will react to that changing its behaviour. After the action of delegation we have in fact a new situation Ω′ (if delegation is the only event that influence the trustworthiness) and we can have two possible results:

i) the new trustworthiness of *B* as for τ is greater than the previous one; at least one of the two possible elementary components is increased: *DoA*, *DoW*; so we can write:

$$\Delta\ trustworthiness\ (B\ \tau)=$$

$$F(\,DoA_{B,\tau\ \Omega'},\ DoW_{B,\tau\ \Omega'}) - F(\,DoA_{B.\tau\ \Omega},\ DoW_{B,\tau\ \Omega}) > 0$$

ii) the new *B*'s reliability as for τ has reduced

$$\Delta\ trustworthiness\ (B\ \tau) < 0.$$

In case (i) *B* has adopted *A*'s goal, i.e. it is doing τ also in order to let/make *A* achieve its goal *g*. Such adoption of *A*'s goal can be for several possible motives, from instrumental and selfish, to pro-social.

The components' degree can change in different ways: the degree of ability (*DoA*) can increase because *B* could use additional tools, new consulting agents, and so on; the degree of willingness (*DoW*) can increase because *B* could have more attention, intention, and so on (the specific goal changes its level of priority).

In case (ii) *B* on the contrary reacts in a negative (for *A*) way to the discovery of *A*'s reliance and exploitation; for some reason *B* is now less willing or less capable in doing τ. In fact in case (ii) too, the reliability components can be independently affected: first, the degree of ability (*DoA*) can decrease because *B* could be upset about the *A*'s exploitation and the *B's* ability could result compromised; again, the willingness degree (*DoW*) can decrease (*B* will have less intention, attention, etc.).

Notice that in this case the change of the *B*'s reliability is not known by *A*. So, even if *A* has a perfect perception of previous *B*'s trustworthiness (that is our hypothesis), in this new situation -with weak delegation- *A* can have an under or over estimation of *B*'s trustworthiness. In other terms, after the weak delegation (and if there is a change of *B*'s trustworthiness following it) we have:

$$DoT_{A,B,\tau,\Omega} \neq trustworthiness(B\ \tau\ \Omega')$$

Let us show you the flow chart for the weak delegation (Figure 2): in it we can see how, on the basis of the mental ingredients of the two agents, the more or less collaborative behaviours of the trustee could be differently interpreted by the trustier. In the case of the mutual knowledge about the awareness of the weak delegation, the trustier could evaluate and learn if *B* is a spontaneous collaborative agent (with respect that task in that situation) and how much *B* is so collaborative (the value of Δ*x*). In the case in which *A* ignores the *B*'s awareness about the weak delegation, the trustier could evaluate the credibility of its own beliefs (both about the *B*'s trustworthiness and about the *B*'s awareness on the weak delegation) and, if the case, revises them.

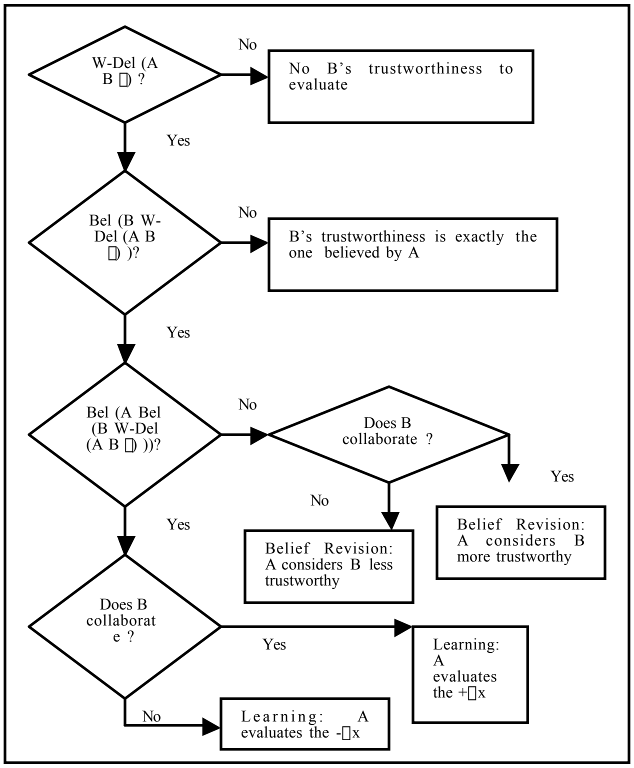
W-Del (A B τ) ?
No
No B's trustworthiness to evaluate
Yes
Bel (B W-Del (A B τ))?
No
B's trustworthiness is exactly the one believed by A
Yes
Bel (A Bel (B W-Del (A B τ)))?
No
Does B collaborate ?
Yes
No
Belief Revision: A considers B more trustworthy
Belief Revision: A considers B less trustworthy
Yes
Does B collaborate ?
Yes
Learning: A evaluates the +Δx
No
Learning: A evaluates the -Δx

Figure 2

In Figure 3 is resumed how weak delegation can influence the delegee's trustworthiness.

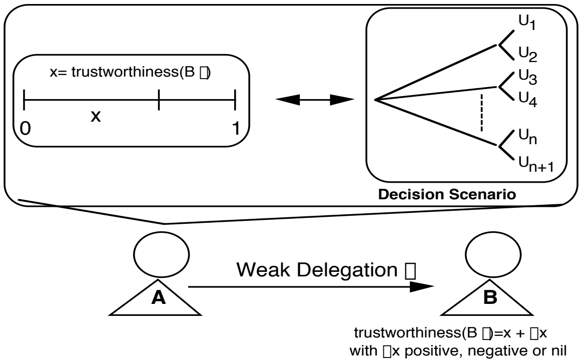

Figure 3

Agent *A* has both a belief about the *B*'s trustworthiness and a hypothetical scenario of the utilities (in the case of

[3] Other possible alternative hypoteses are:

$\neg Bel(A\ Bel(B\ W\text{-}Delegates(A,B,\tau)))\ \wedge\ Bel(B\ W\text{-}Delegates(A,B,\tau))$ or

$Bel(A\ Bel(B\ \neg W\text{-}Delegates(A,B,\tau)))\ \wedge\ Bel(B\ W\text{-}Delegates(A,B,\tau))$

success or failure) of all the possible choices it can do (to delegate to *B* or to *C*, etc., or do not delegate and doing by itself or doing nothing). On this basis it makes a weak delegation and may be it changes the *B*'s trustworthiness. In this last case (changed trustworthiness of the trustee) may be that the A's choice (done before the B's action and of its spontaneous collaboration or of its negative reactions) results better or worst with respect the other previous possibilities.

4.2 The Case of Strong Delegation

We call *strong delegation (S-Delegates(A B □))*, that *based on explicit agreement between A and B*. For a deep analysis on strong delegation see [7].

In this case we have: *MBel(A B S-Delegates(A B □))*

i.e. there is a mutual belief of *A* and *B* about the strong delegation and about the reciprocal awareness of it.

Like in the weak delegation this belief could change the *B*'s trustworthiness, and also in this case we can have two possible results:

i) the new trustworthiness of *B* as for □ is greater than the previous one: □ *trustworthiness(B □)* > 0

ii) the new trustworthiness of *B* on □ is less than the previous one: □ *trustworthiness(B □)* < 0.

Why does *B*'s trustworthiness increase or decrease? In general, a strong delegation increases the trustworthiness of the delegee because of its *commitment*.

This is in fact one of the motives why agents use strong delegation. But it is also possible that the delegee loses motivations when it has to do something not spontaneously but by a contract or by a role.

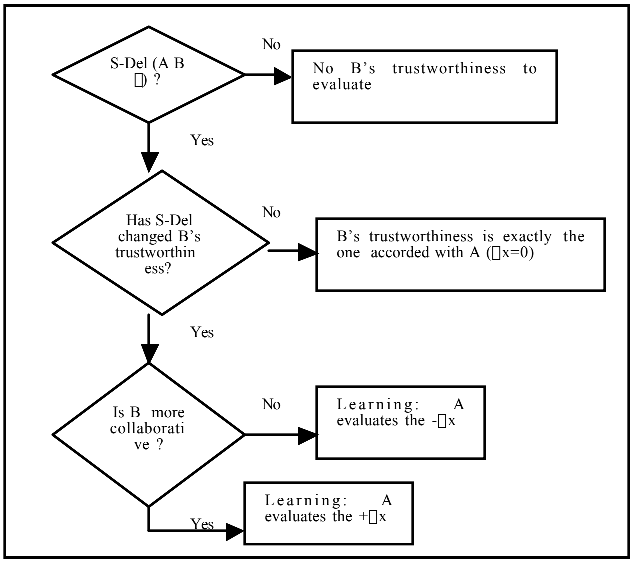
S-Del (A B □) ?
No
No B's trustworthiness to evaluate
Yes
Has S-Del changed B's trustworthiness?
No
B's trustworthiness is exactly the one accorded with A (□x=0)
Yes
Is B more collaborative ?
No
Learning: A evaluates the -□x
Yes
Learning: A evaluates the +□x

Figure 4

The important difference with the previous case is that now *A* knows that *B* will have some possible reactions to the delegation and consequently *A* is expecting a new *B*'s trustworthiness (Fig. 4): $DoT_{A,B,□}$*=trustworthiness(B,□)*

4.3 Anticipated Effects

This is the case in which the delegating agent *A* takes into account the possible effects of its strong delegation on the *B*'s trustworthiness before it performs the delegation action itself: in this way *A* changes the $DoT_{AB□}$ before the delegation (Figure 5).

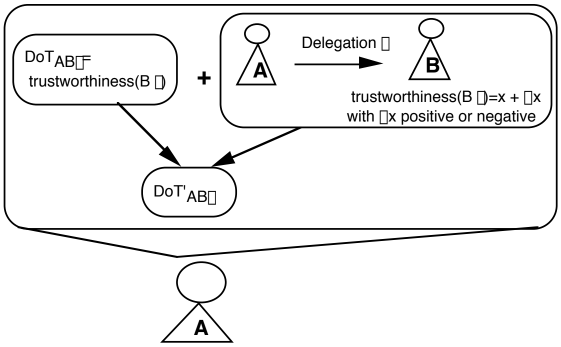

Figure 5

We could have two main interesting subcases:

i) the new degree of trust ($DoT'_{AB□}$) is greater than the old one ($DoT_{AB□}$): $DoT'_{AB□} > DoT_{AB□}$;

ii) it is lesser than the old one: $DoT'_{A\,B\,□} < DoT_{A\,B\,□}$.

In Table 2 are considered all the possible decisions of *A*.

We have called □ the minimum threshold to delegate. In other words, before to perform a delegation action and just for delegating, an agent *A* could evaluate the influence (positive or negative) that its delegation will have on the *B*'s trustworthiness. After this *A* makes the decision. Very interesting is also to evaluate which are the different resulting situations after the trustier's decision to delegate (three cases in Table 2).

	DoT > □ DoT - □ = □ >0 (A would delegate B before evaluating the effects of delegation itself)	DoT < □ DoT - □ = □ <0 (A would not delegate B before evaluating the effects of delegation itself)
DoT'-DoT = □>0 (A thinks that delegation would increase B's trustworthiness)	DoT' - □ = □+ □ >0 **Decision to delegate**	DoT' - □ = □- □ >0 (□> □) **Decision to delegate**
		DoT' - □ = □- □ <0 (□< □) **Decision not to delegate**
DoT'-DoT = □<0 (A thinks that delegation would decrease B's trustworthiness)	DoT' - □ = - □+ □ >0 (□>□) **Decision to delegate**	DoT' - □ = -□- □ <0 **Decision not to delegate**
	DoT' - □ = - □+ □ <0 (□<□) **Decision not to delegate**	

Table 2

In Table 3 we analyze these three cases deriving from the A's decision to delegate.

Even in the case in which the trustee collaborates with the trustier, may happens that the delegated task is not achieved; for example because the expected additional motivation and/or abilities resulting from the delegation act are less effective than the trustier believed.

Trustee's mind Trustier's mind	Collaboration	No collaboration
□ + □>0 □ >0, □>0	□x > 0 □x > □ B more trustworthy □x = □ B equal trustworthy □x < □ B less trustworthy	□x= 0 OR □x < 0 □x = 0 B less trustworthy □x<0 ⟨ |□x|<□ B less trustworthy |□x|>□ **B no trustworthy**
□ + □>0 □ >0, □<0	□x= 0 OR □x > 0 B more trustworthy	□x < 0 |□x|>□ ⟨ |□x|>□ **B no trustworthy** |□x|=□ B less trustworthy |□x|<□ B less trustworthy |□x|=□ B equal trustworthy |□x|<□ B more trustworthy
□ + □>0 □ <0, □>0	□x > 0 □x > □ B more trustworthy □x = □ B equal trustworthy □x<□ ⟨ □x<□+□ **B no tr.** □x>□+□ B less tr.	□x= 0 OR □x < 0 **B no trustworthy**

Table 3

Another interesting way for increasing trustworthiness is through the self-confidence dimension, that we did not explicitly mention since it is part of the ability dimension. In fact, at least in human agents the ability to do □ is not only based on skills (an action repertoire) or on knowing how (library of recipes, etc.), it also requires self-confidence that means the subjective awareness of have those skills and expertise, plus a general good evaluation (and feeling) of its own capability of success. Now the problem is that self-confidence is socially influenced, i.e. my confidence and trust in you can increase your self-confidence. So, I could strategically rely on you (letting you know that I'm relying on you) in order to increase your self-confidence and then my trust in you as for your ability and trustworthiness to do.

5. Concluding Remarks

Strategies and devices for trust building in MAS and virtual societies should take into account the fact that social trust is a very dynamic phenomenon both in the mind of the agents and in the society. We have considered two main basic aspects of this phenomenon:

i) the traditional problem of the trust reinforcement on the basis of the previous (positive or negative) experiences;

ii) the fact that in the same situation *trust is influenced by trust* in several rather complex ways.

With respect the point (i), we have shown how the schema of Figure 6 is not necessarily true.

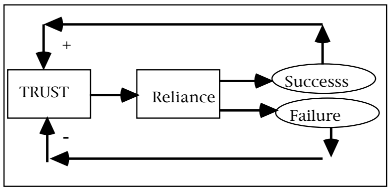

Figure 6

Trust dynamics is more complex than expected and trusting agents may create interference for trust itself.

6. References

[1] C. Jonker and J. Treur (1999), Formal analysis of models for the dynamics of trust based on experiences, *AA'99 Workshop on "Deception, Fraud and Trust in Agent Societies",* Seattle, USA, May 1, pp.81-94.

[2] A. Birk, (2000), Learning to trust, *Autonomous Agents 2000 Workshop on "Deception, Fraud and Trust in Agent Societies",* Barcelona, Spain, June 4, pp.27-38.

[3] S. Barber and J. Kim, (2000), Belief Revision Process based on trust: agents evaluating reputation of information sources, *AA'00 Workshop on "Deception, Fraud and Trust in Agent Societies",* Spain, pp.15-26.

[4] Falcone R., Castelfranchi C. (2001), The socio-cognitive dynamics of trust: does trust create trust? In *Trust in Cyber-societies* R. Falcone, M. Singh, and Y. Tan (Eds.), LNAI 2246 Springer. pp. 55-72.

[5] Castelfranchi C., Falcone R., (1998) Principles of trust for MAS: cognitive anatomy, social importance, and quantification, *Proc. of the Intern. Conference on Multi-Agent Systems (ICMAS'98)*, Paris, July, pp.72-79.

[6] R. Falcone and C. Castelfranchi, (2001), Social Trust: A Cognitive Approach, in C. Castelfranchi and Y. Tan (Eds), Trust and Deception in Virtual Societies, Kluwer Academic Publishers, pp.55-90.

[7] Castelfranchi, C., Falcone, R., (1998) Towards a Theory of Delegation for Agent-based Systems, *Robotics and Autonomous Systems*, SI on Multi-Agent Rationality, Elsevier Editor, Vol 24, Nos 3-4, pp.141-157.

[8] Bratman, (1987), M., E., Intentions, Plans and Practical Reason. Harward University Press, Cambridge Massachusets.

[9] D. Gambetta, editor. Trust. Basil Blackwell, Oxford, 1990.

[10] Castelfranchi, C., Falcone R., Pezzulo, (2003) Trust in Information Sources as a Source for Trust: A Fuzzy Approach, Proc. of AAMAS-03, Melburne, pp.89-96.

[11] Weiner, N., (1961), Cybernetics or Control and Communication in the Animal and the Machine, The MIT Press, Cambridge (Mass.), Wiley and Sons, New York.