6

Dynamics of Trust

Trust in its intrinsic nature is a dynamic phenomenon. Trust has to change on time, because all the entities participating in the trust relationship are potentially modifiable. In real interactions, we never have exactly the same interactive situation in different time instants.

Trust changes with experience, with the modification of the different sources it is based on, with the emotional or rational state of the trustor, with the modification of the environment in which the trustee is supposed to perform, and so on. In other words, being trusted is an attitude depending on dynamic phenomena, as a consequence it is itself a dynamic entity.

In fact, trust is in part a socially emergent phenomenon; it is a mental state, but in socially situated agents and based on social context. In particular, trust is a very dynamic phenomenon; not only because it is based on *the trustor*'s previous experiences, but because it is not simply an external observer's prediction or expectation about a matter of fact.

There are many studies in literature dealing with the dynamics of trust ((Jonker and Treur, 1999), (Barber and Kim, 2000), (Birk, 2000), (Falcone and Castelfranchi, 2001)). We are interested in analyzing four main basic aspects of this phenomenon:

i) The traditional problem of how trust changes on the basis of the trustor's experiences (both positive and negative ones).
ii) The fact that in the same situation *trust is influenced by trust* in several rather complex ways.
iii) How diffuse trust diffuses trust (*trust atmosphere*), that is how X's trusting Y can influence Z trusting Y or W, and so on.
iv) The fact that it is possible to predict how/when an agent who trusts something/someone will therefore trust something/someone else, before and without a direct experience (*trust through generalization reasoning*).

The first case (i) considers the well known phenomenon about the fact that trust evolves in time and has a history, that is X's trust in Y depends on X's previous experience and learning with Y himself or with other (similar) entities. In the following sections, we will analyze this case showing that it is true that in general a successful performance of Y increases X's trust in him (and vice versa a failing performance drastically decreases X's trust) but we will also

Trust Theory: A Socio-Cognitive and Computational Model Cristiano Castelfranchi and Rino Falcone
 ISBN: 978-0-470-02875-9

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).

The dynamic nature of trust is also described by the second case (ii) where we will study the fact that in one and the same situation *trust is influenced by trust* in several rather complex ways. In particular, we will analyze two main crucial aspects of trust dynamics.

How trust creates a reciprocal trust, and distrust elicits distrust; but also vice versa: how *X*'s trust in *Y* could induce lack of trust or distrust in *Y* towards *X*, while *X*'s diffidence can make *Y* more trustful in *X*. In this chapter we will examine also an interesting aspect of trust dynamics: *How the fact that X trusts Y and relies on him in situation* Ω *can actually (objectively) influence Y'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 *X* can be aware of (and takes into account) the effect of its own decision in the very moment of that decision (see also Section 8.9). We will also analyze the trust atmosphere. This is a macro-level phenomenon, and the individual agent does not calculate it. Finally, we will consider the power of the trust cognitive model for analyzing and modeling the crucial phenomenon of a trust transfer from one agent to another or from one task to another.

As we have argued in Chapters 2 and 3, we will resume in the following, that trust and reliance/delegation are strictly connected phenomena: trust could be considered as the set of mental components on which a delegation action is based. In the analysis of trust dynamic, we have also to consider the role of delegation (*weak*, *mild* and *strong* delegation) (Castelfranchi and Falcone, 1998).

6.1 Mental Ingredients in Trust Dynamics

From the point of view of the dynamic studies of trust, it is relevant to underline how the basic beliefs, described in Chapter 2, might change during the interaction between the trustor and the trustee: for example, they could change the abilities of the trustee or his reasons/motives for willing (and/or the trustor's beliefs on them); or again it might change the dependence relationships between the trustor and the trustee (and so on).

Another important characteristic of the socio-cognitive model of trust is the distinction (see also Section 2.7.2 and Section 8.3.3) between trust 'in' someone or something that on the basis of its *internal characteristics* can realize a useful action or performance, and the global trust in the global event or process and its result which is also affected by *external factors* (to the trustee) like opportunities and interferences.

Trust in *Y* (for example, 'social trust' in the strict sense) seems to consist of 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/motivation* (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.

Let us now introduce some formal constructs. We define $Act=\{\alpha_1, \ldots, \alpha_n\}$ 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, beliefs, motives, etc.

As introduced in Chapter 2, the action/goal pair $\tau=(\alpha,g)$ is the real object of delegation, and we called it 'task'.[1] Then by means of τ, we will refer to the action (α), to its resulting world state (g)[2] , or to both.

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

- An *objective degree of ability* (*OdA*, ranging between 0 and 1, indicating the level of Y's ability about the task τ); we can say that it could be measured as the number of Y's successes (s) on the number of Y's attempts (a): *s/a*, when a goes to ∞:

$$OdA_Y = lim\ (a{-}{>}\infty)\ s/a \tag{6.1}$$

and

- An *objective degree of willingness* (*OdW*, ranging between 0 and 1, indicating the level of Y's intentionality/persistence about the task τ); we can say that it could be measured as the number of Y's (successfully or unsuccessfully) performances (p) of that given task on the number of times Y declares to have the intention (i) to perform that task: *p/i*, when i goes to ∞:

$$OdW_Y = lim\ (i{-}{>}\infty)\ p/i \tag{6.2}$$

we are considering that an agent declares its intention each time it has got one.
So, in this model we have that:

$$Trustworthiness\ (Y,\ \tau,\ \Omega) = F(OdA_{Y\tau\Omega},\ OdW_{Y\tau\Omega}) \tag{6.3}$$

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 Y's trustworthiness; for example, if *trustworthiness($Y\,\tau\,\Omega$) = 0.80*, we suppose that in a context Ω, *80%* of times Y tries and succeeds in executing τ.

As the reader can see, we have considered the opportunity dimension as included in Ω: the external conditions favoring, allowing or inhibiting, impeding the realization of the task.

---

[1] 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.*

[2] We consider $g = g_X = p$ (see Chapter 2, Section 2.1).

6.2 Experience As an Interpretation Process: Causal Attribution for Trust

It is commonly accepted ((Jonker and Treur, 1999), (Barber and Kim, 2000), (Birk, 2000)) and discussed in another work of ours (Falcone and Castelfranchi, 2001)) that one of the main sources of trust is direct experience. It is generally supposed that, on the basis of the realized experiences, to each success of the trustee (believed by the trustor) there is a significant increment or a confirmation of the amount of the trustor's trust towards him, and that to every trustee's failure (believed by the trustor) there is a corresponding reduction of the trustor'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 of the trustworthiness's value 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 systems, since it is unable to discriminate cases and reasons of failure and success adaptively. However, this primitive view cannot be avoided until trust is modeled just as a simple index, a dimension, an all-inclusive 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 *X*'s reliance on *Y* and of *Y*'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 *Y*'s failure or success on *X*'s trust in *Y* depends on *X*'s *'causal attribution'* ((Weiner, 1992)) 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 *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*.

Is Yody's guilt or merit based on whether he was failing or successful on τ? Or was the real responsibility about the conditions in which he worked? Was his performance the standard performance he was able to realize? Were the environmental conditions the standard ones in which that task is realized?

The cognitive, emotional, and practical consequences of a failure (or success) *strictly depend 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 (see Chapter 9); 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 6.1).

Since in agent-mediated human interaction (like Computer Supported Cooperative Work or Electronic Commerce) and in cooperating autonomous Multi-Agent Systems 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. However, this holds also for marketing and its model of the consumer's trust and loyalty towards a brand or a shop or a product/service; and for trust dynamics in interpersonal relations; and so on.

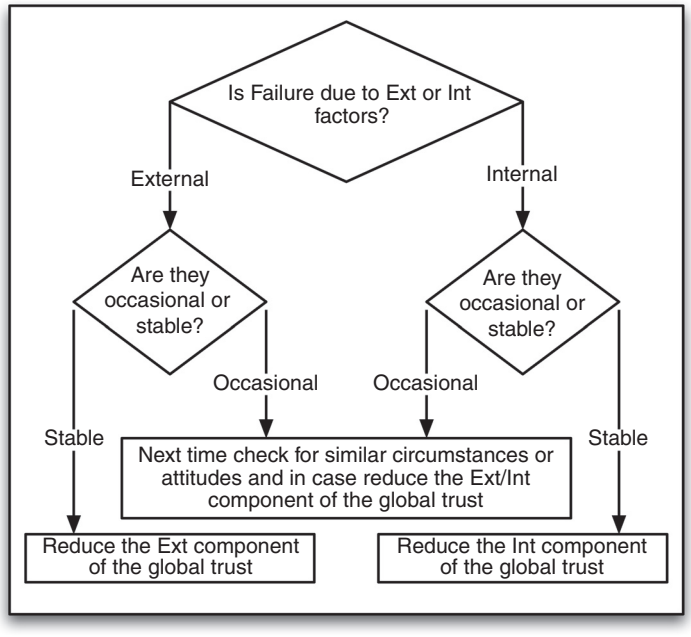

Figure 6.1 Flow-chart to identify the causes of failure/success

Let's first present a basic model (which exploits our cognitive analysis of trust attitude and the '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 X evaluates its own trust (*degree of trust*) in agent Y about the task τ (to be performed) in the environment Ω ($DoT_{X,Y,\tau,\Omega}$):

$$DoT_{X,Y,\tau,\Omega} = f(DoA_{X,Y,\tau},\ DoW_{X,Y,\tau},\ e(\Omega)) \tag{6.4}$$

Where: f (like F) is a general function that preserves monotonicity. In particular, $DoA_{X,Y,\tau}$ is Y's degree of ability (in X's opinion) about the task τ; $DoW_{X,Y,\tau}$ is Y's degree of motivational disposition (in X's opinion) about the task τ (both $DoA_{X,Y,\tau}$ and $DoW_{X,Y,\tau}$ are evaluated in the case in which Y 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 Y (this part cannot be considered as a separated task but as an 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 *($DoA_{X,Y,\tau}$ (Abilities), $DoW_{X,Y,\tau}$ (Motivations),* and *e(Ω) (Environment))* of X'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.
   We assume this for the 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 delivering a piece of important machinery 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* on each other.
We have two sub-cases: first, the *Environment* favours or disfavours the *Y*'s *Abilities* (useful for the task achievement); second, the *Y*'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 case (iii), there are two sub-cases: first, the *Environment* influences *Y*'s *Motivations* (useful for the task achievement); second, *Y*'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, a well informed trustor who is supplied with an analytic apparatus (a socio-cognitive agent), could evaluate which ingredients performed well and which failed in each specific experiential event (analyzing and understanding the different role played by each ingredient in the specific performance).

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 trustor could verify:

$$Actual(DoA,\ DoW) - Expected(DoA,\ DoW) > 0 \tag{6.5}$$

$$Actual(DoA,\ DoW) - Expected(DoA,\ DoW) < 0 \tag{6.6}$$

$$Actual(e(\Omega)) - Expected(e(\Omega)) > 0 \tag{6.7}$$

$$Actual(e(\Omega)) - Expected(e(\Omega)) < 0 \tag{6.8}$$

Where the operators *Actual* and *Expected* give the values of the arguments as respective evaluations after the performance of the task and before it.

In (6.5) and (6.7) both the trustee (*internal-trust*) and the environment (*external-trust*) are more trustworthy than expected by the trustor; vice versa, in (6.6) and (6.8) they are both less trustworthy than expected by the trustor.

In Table 6.1 all the possible combinations are shown.

Where: 'More Int-trust' ('Less Int-trust') means that the trustor after the performance considers the trustee more (less) trustworthy than before it (he performed better (worst) than expected); 'More Ext-trust' ('Less Ext-trust') means that the trustor after the performance considers the environment more (less) trustworthy than before it (it performed better (worst) than expected).

Table 6.1 Performances of the trustee and the environment in combination with the success or failure of the global task

	Success of the performance	Failure of the performance
Δ**(int-trust)** $>$ **0**	**A**	**A′**
Δ**(ext-trust)** $>$ **0**	More Int-trust; More ext-trust	More Int-trust; More ext-trust
Δ**(int-trust)** $>$ **0**	**B**	**B′**
Δ**(ext-trust)** $<$ **0**	More Int-trust; Less ext-trust	More Int-trust; Less ext-trust
Δ**(int-trust)** $<$ **0**	**C**	**C′**
Δ**(ext-trust)** $>$ **0**	Less Int-trust; More ext-trust	Less Int-trust; More ext-trust
Δ**(int-trust)** $<$ **0**	**D**	**D′**
Δ**(ext-trust)** $<$ **0**	Less Int-trust; Less ext-trust	Less Int-trust; Less ext-trust

Cases of particular interest are:

(*B in Table 6.1*) in which even if the environment is less trustworthy than expected, the better performance of the trustee produces a global success performance.

In fact, three factors have to be considered: the trustor over-evaluated the environmental's trustworthiness; she under-evaluated the trustee's trustworthiness; the composition of internal and external factors produced a successful performance.

(*C in Table 6.1*) in which even if the trustee is less trustworthy than expected, the better performance of the environment produces a global success performance.

Also in this case, three factors have to be considered: the trustor under-evaluated the environmental's trustworthiness; she over-evaluated the trustee's trustworthiness; the composition of internal and external factors produced a successful performance. Tom actually made a mess, was really a disater, but was also incredibly 'lucky': accidentally and by circumstantial factors eventually the desired result was there. But not thanks to his ability or willingness!

This is a very interesting case in which the right causal attributions make it possible the trust in the trustee to decrease even in presence of his success. Of course, two main possible consequences follow: the new attributed trustworthiness of the trustee is again (in the trustor's view) sufficiently high (over a certain threshold) for trusting him later; vice versa, the new attributed trustee's trustworthiness is now unsufficient for trusting him later (under the threshold).

(*D and A' in Table 6.1*) In which expectations do not correspond with the real trustworthiness necessary for the task: too high (both in the trustor and in the environment) in D and too low (both in the trustor and in the environment) in A'. These cases are not possible if the trustor has a good perception of the necessary levels of trustworthiness for that task (as we suppose in the other cases in Table 6.1).

(*B' in Table 6.1*) in which even if the trustee is more trustworthy than expected (so increases the trust in him), the unexpected (at least for the trustor) difficulties introduced by the environment produce a failure of the global performance. *This is another interesting case in*

which the right causal attributions make it possible to increase the trust in the trustee even in the presence of his failure.

Consider this situation: 'In the last tour in Italy I saw that the coach driver did his best to arrive on time at the Opera in Milano, and was very competent and committed: to recognise his shortcomings, drive quickly, but safely, and do everything he could, but unfortunately, given that we were already late because of some of the passengers and due to the traffic on the highways we missed the Opera. In any case my trust in that driver actually increased due to the respect I had in him and I would definitely use him again'.

Again, the case is made more complex when there is some dependence between the internal properties and the environment (cases (iii) and (iv). In this case, in addition to the introduced factors Δ*(int-trust)* and Δ*(ext-trust)*, we have to also take into account the possible influences between internal and external factors. We consider these influences as not expected from the trustor in the sense that the expected influences are integrated directly in the internal or external factors. We can – for example – consider the case of a violinist. We generally trust him for playing very well; but, suppose he has to do the concert in an open environment and the weather conditions are particularly bad (very cold): may be these conditions can modify the specific hand abilities of the violinist and his performance; at the same time, it is possible that a special distracted, inattentive, noisy audience could modify his willingness and consequently again his performance.

Concluding with the experience-based trust we have to say that the important thing is not only the final result of the trustee's performance but in particular *the trustor's causal attribution to all the factors producing that result*. It is on the basis of these causal attributions that the trustor updates her beliefs about the trustworthiness of the trustee, of the environment, and of their reciprocal influences.

So the rational scheme is not the simplified one showed in Figure 6.2 (where there is a trivial positive or negative feedback to Y on the basis of the global success or global failure), but the more complex one showed in Figure 6.3 (where in the case of either failure or success both the components and their specific contributions are considered).

6.3 Changing the Trustee's Trustworthiness

In this paragraph we are going to analyze how a delegation act (corresponding to a decision making based on trust in a specific situational context) could change (just because of that delegation action and in reaction to it) the trustworthiness of the delegated agent (*delegee*). This not only holds in *Strong-Delegation* (where Y is aware of it and X counts on his awareness, adhesion, and commitment, and there is some explicit or implicit communication of X's trust and reliance; Section 2.6.1). It even holds in *Weak-Delegation* (where Y and his autonomous behavior – in X's intention – is simply exploited by X), but in peculiar conditions.

6.3.1 The Case of Weak Delegation

As also shown in Section 2.6, we call the reliance simply based on exploitation for the achievement of the task *weak delegation* (and express this with *W-Delegates*$(X\ Y\ \tau)$). In it there is no agreement, no request or even (intended) influence: *X is just exploiting in its plan*

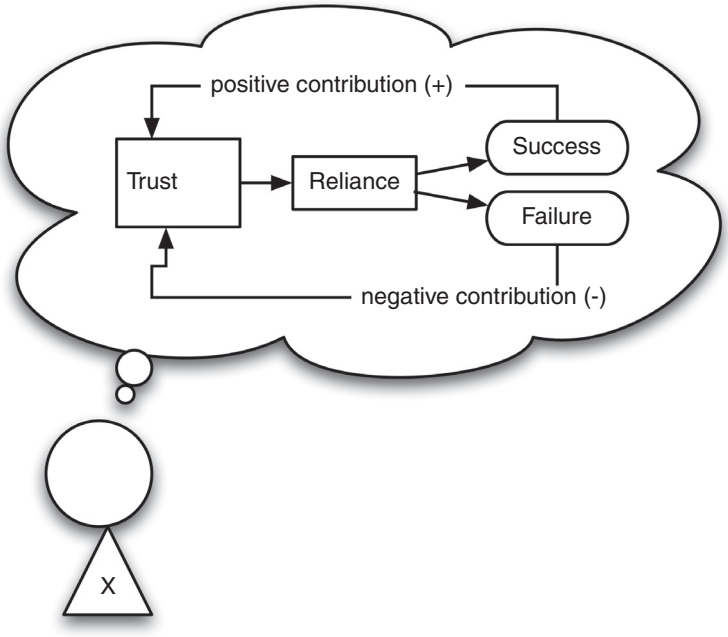

Figure 6.2 Simplified scheme of the performances' influences on Trust Evaluation

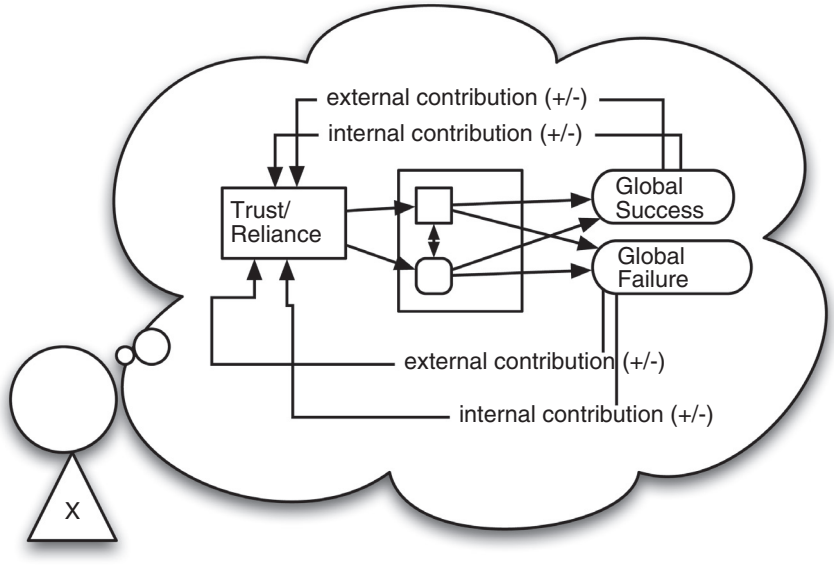

Figure 6.3 Realistic (with causal attribution) scheme of the performances' influences on Trust Evaluation

a fully autonomous action of Y. For a more complete discussion on the mental ingredients of the weak delegation see (Castelfranchi and Falcone, 1998).

The expression *W-Delegates(X Y τ)* represents the following *necessary* mental ingredients:

a) The achievement of τ (the execution of α and its result g) is a *goal* of X.
b) X believes that there exists another agent Y that has the *power of* (Castelfranchi, 1991) achieving τ.
c) X believes that Y will achieve τ in time and by itself (without any X's intervention).
   c-bis) X believes that Y *intends* (in the case that Y is a cognitive agent) to achieve τ in time and by itself, and that will do this in time and without any intervention of X.
d) X *prefers*[3] to achieve τ through Y.
e) The achievement of τ through Y is the choice (goal) of X.
f) X has the goal (*relativized* (Cohen and Levesque, 1987) to (e)) of not achieving τ by itself.

We consider (*a*, *b*, *c*, and *d*) what the agent X views as a '*Potential for relying on*' the agent Y, its *trust* in Y; and (*e* and *f*) what X views as the '*Decision to rely on*' Y. We consider 'Potential for relying on' and 'Decision to rely on' as two constructs temporally and logically related to each other.

We hypothesize that in *weak delegation* (as in any delegation) there needs to be a decision made based on trust and in particular there are two specific beliefs of X:

$belief1_X$: if Y makes the action then Y has a successful performance;

$belief2_X$: Y intends to do the action.

For example, X sees Y waiting at the bus stop, and – while running to catch the bus – she counts on Y to stop it.

As shown in Section 6.2, the trustworthiness of Y is evaluated by X using the formula (6.4). For the sake of simplicity we assume that:

$$DoT_{X,Y,\tau,\Omega} \equiv trustworthiness\ (Y\ \tau\ \Omega) \tag{6.9}$$

In other words: X has a perfect perception of Y's trustworthiness: X believes/knows Y's real trustworthiness.

The interesting stage in weak delegation is when:

$$Bel(X \neg Bel(Y\ W\text{-}Delegates\ (X,\ Y,\ \tau))) \cap Bel(Y\ W\text{-}Delegates(X, Y, \tau))^4 \tag{6.10}$$

in other words: there is a weak delegation by X on Y, but actually Y is aware of it (while X believes that Y is not).

The first belief ($belief1_X$) is very often true in weak delegation, while the second one ($belief2_X$) is necessary in the case we are going to consider. If *Bel(Y W-Delegates(X Y τ))*, this

[3] This means that, either relative to the achievement of τ or relative to a broader goal g' that includes the achievement of τ, X believes herself to be dependent on Y (see (Jennings, 1993), (Sichman *et al.*, 1994) and Section 2.9.1 on 'weak dependence').

[4] Other possible alternative hypoteses are:
$\neg Bel(X\ Bel(Y\ W\text{-}Delegates(X,\ Y, \tau))) \cap Bel(Y\ W\text{-}Delegates(X, Y, \tau))$ *or*
$Bel(X\ Bel(Y \neg W\text{-}Delegates(X, Y, \tau))) \cap Bel(Y\ W\text{-}Delegates(X, Y, \tau))$

belief could change *Y*'s trustworthiness, either because *Y* will adopt *X*'s goal as an additional motivation and accept such an exploitation, or because, on the contrary, *Y* will refuse such an exploitation, changing his behaviour and reacting to the delegation (there is in fact also a third case, in which this knowledge does not influence *Y*'s behaviour and beliefs: we do not consider this case). After the action of delegation we have in fact a new situation Ω' (if delegation is the only event that influences the trustworthiness) and we can have two possible results:

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

$$\Delta trustworthiness(Y\ \tau) = F(OdA_{Y,\tau,\Omega'},\ OdW_{Y,\tau,\Omega'})\text{-}F(OdA_{Y,\tau,\Omega}, OdW_{Y,\tau,\Omega}) > 0 \tag{6.11}$$

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

$$\Delta trustworthiness\ (Y\ \tau) < 0 \tag{6.12}$$

In case (*6.11*) *Y* has adopted *X*'s goal, i.e. he is doing τ also in order to let/make *X* achieve its goal *g*. Such adoption of *X*'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 (*OdA*) can increase because, for example, *Y* could invest more attention in the performance, use additional tools, new consulting agents, and so on; the degree of willingness (*OdW*) can increase because *Y* could have additional motives and a firm intention, and so on (the specific goal changes its level of priority).

In case (*6.12*) *Y* on the contrary reacts in a negative way (for *X*) to the discovery of *X*'s reliance and exploitation; for some reason *Y* is now less willing or less capable of doing τ. In fact in case (ii) too, the reliability components can be independently affected: first, the degree of ability (*OdA*) can decrease because *Y* could be upset about the *X*'s exploitation and *Y's* ability could be compromised; again, the willingness degree (*OdW*) can decrease (*Y* will have less intention, attention, etc.).

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

$$DoT_{X,Y,\tau,\Omega} \neq trustworthiness(Y\ \tau\ \Omega') \tag{6.13}$$

Let us show you the flow chart for the weak delegation (Figure 6.4): 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 trustor. In the case of the mutual knowledge about the awareness of the weak-delegation (but not interpreted as a tacit request and agreement), the trustor could evaluate and learn if *Y* is a spontaneous collaborative agent (with respect that task in that situation) and how much *Y* is so collaborative (the value of Δ). In the case in which *X* ignores *Y*'s awareness about the weak delegation, the trustor could evaluate the credibility of its own beliefs (both about *Y*'s trustworthiness and about *Y*'s awareness of the weak delegation) and, if the case, revises them.

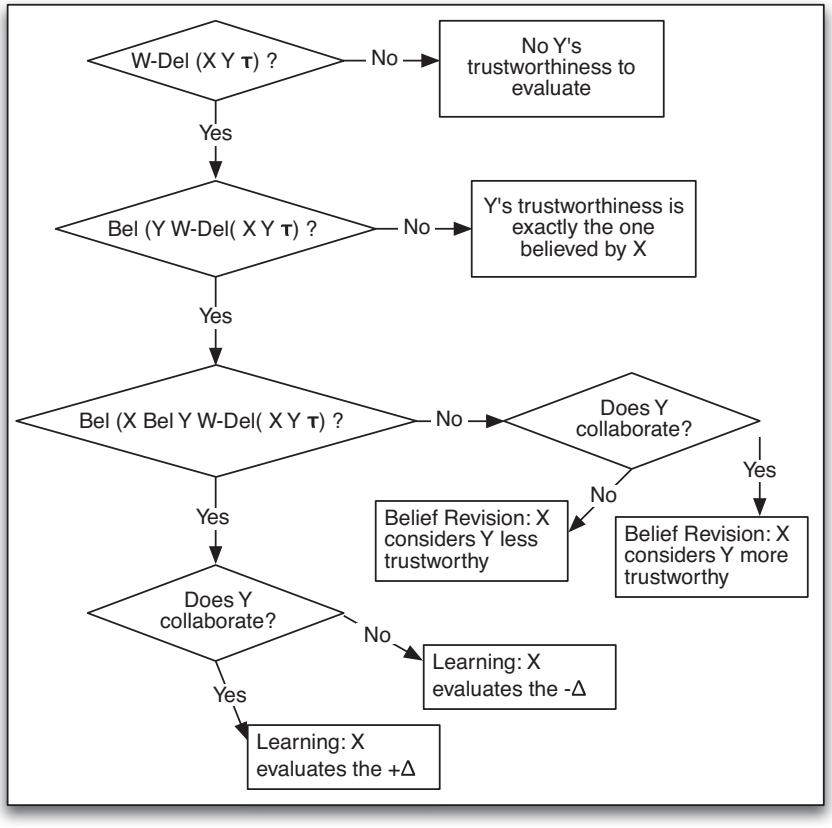

Figure 6.4 Flow-Chart resuming the different mental situations in weak-delegation

In Figure 6.5, it reiterated how weak delegation can influence the delegee's trustworthiness. Agent *X* has both a belief about *Y*'s trustworthiness and a hypothetical scenario of the utilities (in the case of success or failure) of all the possible choices it can do (to delegate to *Y* or to *W*, etc., or to not delegate and do it on its own or do nothing). On this basis it makes a weak delegation and maybe it changes *Y*'s trustworthiness. In the last case (changed trustworthiness of the trustee) maybe that *X*'s choice (done before *Y*'s action and of its spontaneous collaboration or of its negative reactions) results better or worst with respect to the previous possibilities. In other words, in the case in which the weak delegation changes *Y*'s trustworthiness (without *X* being able to foresee this change), the new trustworthiness of *Y* will be different from the expected one by *X* (and planned in a different decision scenario).

6.3.2 The Case of Strong Delegation

We call *strong delegation (S-Delegates(X Y τ))*, that *based on (explicit) agreement between X and Y.*

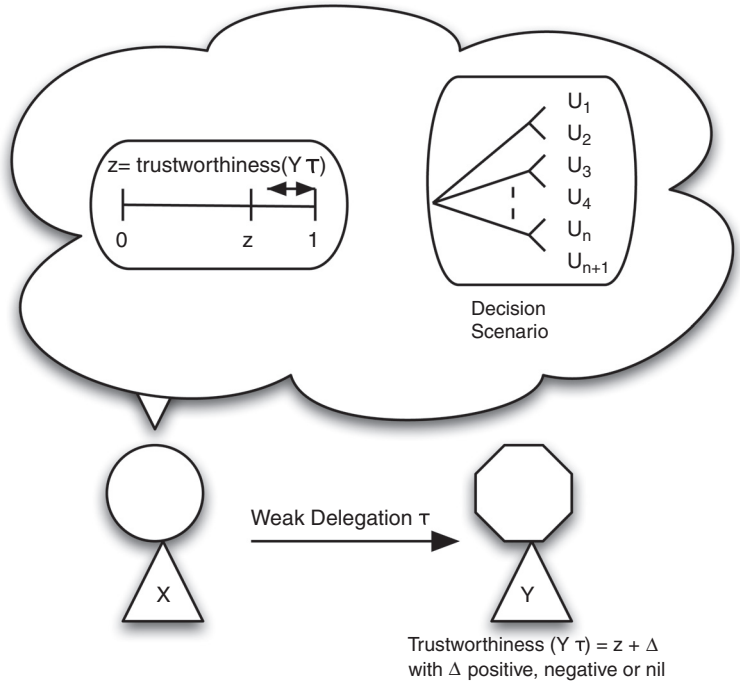

Figure 6.5 Potential Influence of the Weak-Delegation on the Decision Scenario. (Reproduced with kind permission of Springer Science+Business Media © 2001)

The expression *S-Delegates(X Y τ)* represents the following *necessary* mental ingredients:

a′) The achievement of τ is a *goal* of *X*.

b′) *X* believes that there exists another agent *Y* that has the *power of* achieving τ.

c′) *X* does not believe that *Y* will achieve τ by itself (without any intervention of *X*).

d′) *X* believes that if *X* realizes an action α' there will be this result: *Y* will intend τ as the consequence of the fact that *Y* adopts *X's* goal that *Y* would intend τ (in other words, *Y* will be socially committed with *X*).

e′) *X prefers* to achieve τ through *Y*.

f′) *X* intends to do α' relativized to (d').

g′) The achievement of τ through *Y* is the goal of *X*.

h′) *X* has the goal (*relativized* to (g')) of not achieving τ by itself.

We consider (a', b', c', d' and e') what the agent *X* views as a '*Potential for relying on*' the agent *Y*; and (f', g' and h') what *X* views as the '*Decision to rely on*' *Y*.

In this case we have: *MBel(X Y S-Delegates(X Y τ))*

i.e. there is a mutual belief of X and Y about the strong delegation and about the reciprocal awareness of it.

Like in the weak delegation, this belief could change the trustworthiness of Y, and also in this case we can have two possible results (we exclude also in this case the fact that this action does not have any influence on Y):

i) the new trustworthiness of Y as for τ is greater than the previous; so in this case we have the situation given from the formula (6.11): $\Delta trustworthiness(Y\ \tau) > 0$
ii) the new trustworthiness of Y as for τ is less than the previous one; so in this case we have the situation given from the formula (6.12): $\Delta trustworthiness(Y\ \tau) < 0$.

Why does Y's trustworthiness increase or decrease? In general, a strong delegation – if accepted and complete – increases the trustworthiness of the delegee because of its *commitment*.

This is in fact one of the motives why agents use strong delegation and count on Y's 'adhesion' (Section 2.8). However, it is also possible that the delegee loses motivation when he has to do something not spontaneously but by a contract, or by a role or duty, or for somebody else.

The important difference with the previous case is that now X knows that Y will have some possible reactions to the delegation and consequently X is expecting a new trustworthiness of Y (Figure 6.6): in some measure even if there is an increase in Y's trustworthiness it is not completely unexpected by X.

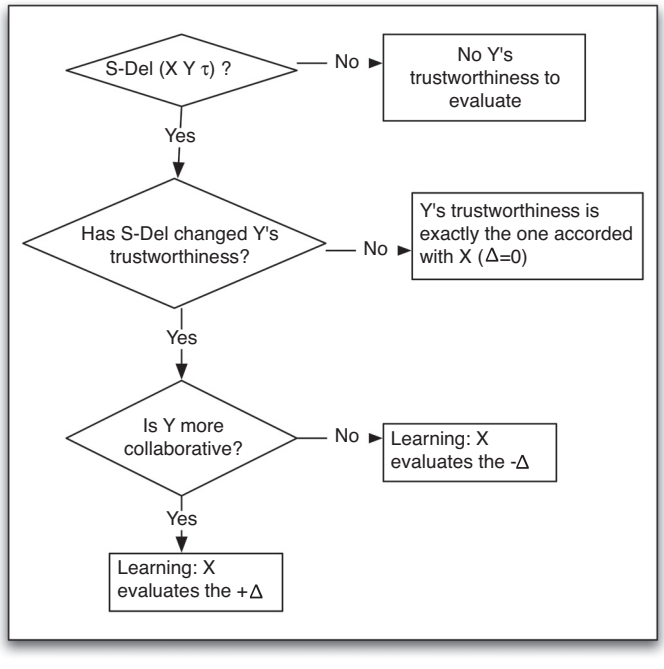

Figure 6.6 Flow-Chart resuming the different mental situations in strong-delegation

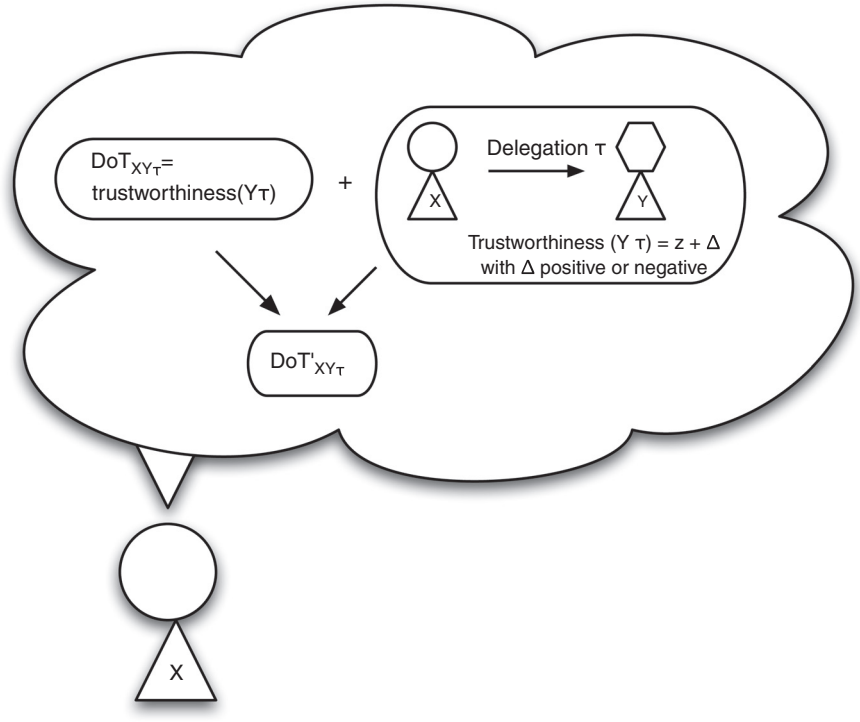

Figure 6.7 Changed Degree of Trust anticipated in strong-delegation. (Reproduced with kind permission of Springer Science+Business Media © 2001)

6.3.3 Anticipated Effects: A Planned Dynamics

This is the case in which the delegating agent *X* takes into account the possible effects of its 'strong delegation' and of the act of trusting (which becomes per se a 'signal'; Section 2.2.7) on *Y*'s trustworthiness, in her decision, before she performs the delegation action. In this way, *X* changes her degree of trust in *Y* ($DoT_{XY\tau}$) before the delegation and on the basis of this her own action (Figure 6.7).

We analyze two cases:

i) the new degree of trust ($DoT'_{XY\tau}$) is greater than the old one ($DoT_{XY\tau}$): $DoT'_{XY\tau} > DoT_{XY\tau}$;
ii) the new degree of trust is lesser than the old one: $DoT'_{XY\tau} < DoT_{XY\tau}$.

Introducing a minimum threshold to delegate (σ) we can analyze the various cases by crossing the change of the *DoT* with the values of the threshold.

In other words, before performing a delegation action, an agent *X* could evaluate the (positive or negative) influence that its delegation will have on *Y*'s trustworthiness and if this influence is relevant for either overcoming or undergoing the minimum level for delegating. In Table 6.2, all the possible situations and the consequent decisions of *X* are considered.

Table 6.2 Crossing different *DoTs* with minimum thresholds to delegate

	$DoT > \sigma$ $DoT - \sigma = \varepsilon' > 0$ (X would delegate Y before evaluating the effects of delegation itself)	$DoT < \sigma$ $DoT - \sigma = \varepsilon' < 0$ (X would not delegate Y before evaluating the effects of delegation itself)
$DoT' - DoT = \varepsilon > 0$ (X thinks that delegating Y would increase Y's trustworthiness)	$DoT' - \sigma = \varepsilon + \varepsilon' > 0$ $(\varepsilon > 0; \varepsilon > 0)$ **Decision to delegate**	$DoT' - \sigma = \varepsilon + \varepsilon' > 0$ $(\varepsilon > 0; \varepsilon' < 0; \varepsilon > \varepsilon')$ **Decision to delegate** $DoT' - \sigma = \varepsilon + \varepsilon' < 0$ $(\varepsilon > 0; \varepsilon' < 0; \varepsilon < \varepsilon')$ **Decision not to delegate**
$DoT' - DoT = \varepsilon < 0$ (X thinks that delegating Y would decrease Y's trustworthiness)	$DoT' - \sigma = \varepsilon + \varepsilon' > 0$ $(\varepsilon < 0; \varepsilon' > 0; \varepsilon < \varepsilon')$ **Decision to delegate** $DoT' - \sigma = \varepsilon + \varepsilon' < 0$ $(\varepsilon < 0; \varepsilon' > 0; \varepsilon > \varepsilon')$ **Decision not to delegate**	$DoT - \sigma = \varepsilon + \varepsilon' < 0$ $(\varepsilon < 0; \varepsilon' < 0)$ **Decision not to delegate**

In Table 6.3, we analyze the three cases deriving from *X*'s decision to delegate as shown in Table 6.2. (In Table 6.3 we call TTE = Trustworthy than expected). Even in the case in which the trustee collaborates with the trustor, it may happen 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 trustor believed.

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 α 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 having those skills and expertise, plus a general good evaluation (and feeling) of our 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.

Finally, another interesting case in strong delegation is when there is the decision to rely upon Y but with diffidence and without any certainty that Y will be able to achieve τ. This is the case in which there is not enough trust but there is delegation (there are a set of possible reasons for this which we do not consider here). We are interested in the case in which Y realizes that anomalous situation (let us call this: diffidence). We have:

Table 6.3 Cases in which there has been delegation to Y

Trustee's Performance / Trustier's Mind	Good Collaboration	Bad Collaboration
$\epsilon + \epsilon' > 0$ ($\epsilon > 0; \epsilon' > 0$)	$\Delta > 0$ $\Delta > \epsilon$; (Y more TTE) $\Delta = \epsilon$; (Y equal TTE) $\Delta < \epsilon$; (Y less TTE)	$\Delta < 0 \cup \Delta = 0$ $\Delta = 0$; (Y less TTE) $\Delta < 0 \cap \lvert\Delta\rvert < \epsilon'$; (Y less TTE) $\Delta < 0 \cap \lvert\Delta\rvert > \epsilon'$; (**Y no trustworthy**)
$\epsilon + \epsilon' > 0$ ($\epsilon < 0; \epsilon' > 0$)	$\Delta > 0 \cup \Delta = 0$ (Y more TTE)	$\Delta < 0$ $\lvert\Delta\rvert > \epsilon \cap \lvert\Delta\rvert > \epsilon'$; (**Y no trustworthy**) $\lvert\Delta\rvert > \epsilon \cap \lvert\Delta\rvert = \epsilon'$;(Y less TTE) $\lvert\Delta\rvert > \epsilon \cap \lvert\Delta\rvert < \epsilon'$; (Y less TTE) $\lvert\Delta\rvert = \epsilon$; (Y equal TTE) $\lvert\Delta\rvert < \epsilon$; (Y more TTE)
$\epsilon + \epsilon' > 0$ ($\epsilon > 0; \epsilon' < 0$)	$\Delta < 0$ $\lvert\Delta\rvert > \epsilon$; (Y more TTE) $\lvert\Delta\rvert = \epsilon$; (Y equal TTE) $\lvert\Delta\rvert < \epsilon \cap \lvert\Delta\rvert < \epsilon + \epsilon'$; (**Y no trustworthy**) $\lvert\Delta\rvert < \epsilon \cap \lvert\Delta\rvert > \epsilon + \epsilon'$; (Y less TTE)	$\Delta < 0 \cup \Delta = 0$ (**Y no trustworthy**)

S-Delegates(X Y τ)
with $DoT_{X\,Y\,\tau} \leq \sigma$; where σ is a minimal 'reasonable threshold' to delegate the task τ and

Bel (Y ($DoT_{X\,Y\,\tau} \leq \sigma$))
Such a diffidence could be implicit or explicit in the delegation: it is not important.

Neither is it important, in this specific analysis, if

$DoT_{X\,Y\,\tau} \neq$ *trustworthiness(Y τ) or* $DoT_{X\,Y\,\tau} =$ *trustworthiness(Y τ)*
(in other words, if X's diffidence in Y is objectively justified or not).

This distrust could, in fact, produce a change (either positive or negative) in *Y*'s trustworthiness:

- *Y* could be disturbed by such a bad evaluation and have a worst performance (we will not consider here, but in general in these cases of diffidence there is always some additional actions by the delegating agent: more control, some parallel additional delegation, and so on);

- *Y*, even if disturbed by the diffidence, could have a pridefully reaction and produce a better performance.

In sum, *Y*'s trustworthiness could be different (with respect to *X*'s expected one) and the cause of this difference could be both *X* 's trust and/or *X* 's distrust.

Let us in particular stress the fact that the predicted effect of the act of trusting *Y* on *Y*'s possible performance can feedback on *X*'s decision, and modify the level of trust and thus the decision itself:

- *X*'s trust might be unsufficient for delegating to *Y*, but the predicted effects of trusting him might make it sufficient!
- Vice versa, *X*'s static trust in *Y* might be sufficient for delegating to him, but the predicted effect of the delegation act on *Y* feeds back on the level of trust, decreases it, and makes it unsufficient for delegating.

6.4 The Dynamics of Reciprocal Trust and Distrust

The act of trusting somebody (i.e. the reliance) can also be an implicitly communicative act. This is especially true when the delegation is *strong* (when it implies and relies on the understanding and agreement of the delegee), and when it is part of a bilateral and possibly reciprocal relation of delegation-help, like in social exchange. In fact, in *social exchange X*'s adoption of *Y*'s goal is *conditional* to *Y*'s adoption of *X*'s goal. *X*'s adoption is based on *X*'s trust in *Y*, and vice versa. Thus, *X*'s trusting *Y* for delegating to *Y* a task τ is in some sense conditional on *Y*'s trusting *X* for delegating to *X* a task τ'. *X* has also to trust (believe) that *Y* will trust her and vice versa: *there is a recursive embedding of trust attitudes.* Not only this, but the measure of *X*'s trusting *Y* depends on, varies with the decision and the act of *Y*'s trusting *X* (and vice versa).

The act of trusting can have among its effects that of determining or increasing Y's trusting X. Thus, *X* may be aware of this effect and may plan to achieve it through her act of trusting. In this case, *X* must plan for *Y* to understand her decision/act of trusting *Y*. But, why does *X* wants to communicate to *Y* about her decision and action of relying on *Y*? In order to induce some (more) trust in *Y*. Thus the higher goal of that communication goal in *X*'s plan is to induce *Y* to believe that '*Y* can trust *X* since *X* trusts *Y*'. And this is eventually in order (to higher goal) to induce *Y* to trust *X*. As claimed in sociology (Gambetta, 1990) there is in social relations the necessity of actively promoting trust. 'The concession of trust' – which generates precisely that behaviour that seems to be its logical presupposition – is part of a strategy for structuring social exchange.

In sum, usually there is a circular relation, and more precisely a positive feedback, between trust in reciprocal delegation-adoption relations (from commerce to friendship). That – in cognitive terms – means that the (communicative) act of trusting and eventually delegating impacts on the beliefs of the other ('trust' in the strict sense) that are the bases of the 'reliance' attitude and decision producing the external act of delegating (Figure 6.8).

Analogously there is a positive feedback relation between distrust dispositions: as usually trust induces trust in the other, so usually distrust increments distrust. What precisely is the

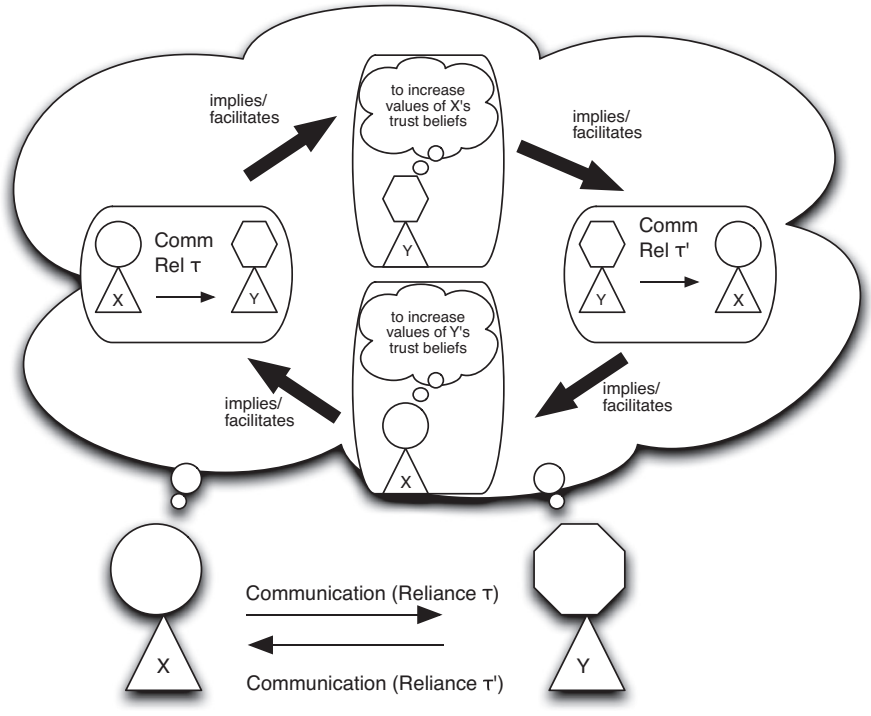

Figure 6.8 Reciprocal Trust and involved mental attitudes

mechanism of trust producing trust in exchange?[5] In my trust about your willingness to exchange (help and delegation) is included my trust that you trust me (and then delegate and then accept to help): if you do not trust me, you will not delegate me and then you will not adopt my goal. And vice versa. Now my trusting you (your believing that I trust you) may increase your trust in me as for delegating for exchange: since I delegate to you, (conditionally, by exchange) you can believe that I am willing to adopt your goal, so you can delegate to me. There are also other means to change *Y*'s willingness to trust us and delegate to us (and then for example exchange with us). Suppose for example that *X* is not sure about *Y* trusting her because of some bad reputation or some prejudice about her. She can change *Y*'s opinion about herself (ex. through some recommendation letter, or showing her good behavior in other interactions) in order to improve *Y*'s trust in *X* and then *Y*'s willingness, and then her own trust in delegating to him.

An interesting concept is the so-called *reciprocal trust* that is not simply *bilateral trust*. It is not sufficient that *X* trusts *Y* and *Y* trusts *X* at the same time. For example, *X* relies on *Y* for

[5] In general, in reciprocal relationship, trust elicits trust. This is also typical in friendship. If he trusts me (in relation to keeping secrets, not cheating, and not making fun of), he cannot have aggressive motives against me (he would expose himself to retaliation; he should feel antipathy, but antipathy does not support confidence). So, he must be benevolent, able to keep secrets, not cheat on me, and not make fun of me: this is my trust in reciprocating confidences.

stopping the bus and *Y* relies on *X* for stopping the bus, there is bilateral (unaware) trust and delegation; nobody stops the bus and both fail!

To have *reciprocal trust*, mutual understanding and communication (at least implicit) are needed: *X* has the goal that *Y* knows that she (will) trust *Y*, in order that *Y* trusts her, and that she trusts *Y* if and only if *Y* trusts her.

Exchange is in fact characterized by *reciprocal conditional trust*: how the act of trusting can increase *Y*'s trust and then my own trust which should be presupposed by my act of trusting. However, no paradox or irrationality is there, since my prediction of the effect anticipates my act and justifies it. For example, *X* can be more sure (trust) about *Y*'s motivation, because she is proposing to *Y* (or accepting from *Y*) a specific and reliable motive of *Y* for doing as expected: an *instrumental* goal-adoption for *selfish reasons* (see Section 2.8 A. Smith citation). *Y* has a specific interest and advantage for doing something 'for' *X* (provided that he believes that *X* will do as expected).

Given our agents *X* and *Y* and two possible tasks: τ and τ', we suppose that:

$$DoT_{XY\tau} < \sigma_1$$ [6]

The value of the $DoT_{XY\tau}$ changes on the basis of its components' variation. One of the ways to change the elementary components of $DoT_{XY\tau}$ is when the trustee (*Y* in our case) communicates (explicitly or implicitly) his own trust in the trustor (*X*) as for another possible task (τ') for example delegating the task τ' to *X* (relying upon *X* as for the task τ'). In our terms will be true the formula:

S-Delegation(*Y X* τ') (that always implies *MutualBel* (*X Y S-Delegation*(*Y X* τ'))).

In fact, this belief has various possible consequences in *X*'s mind:

i) there exists a dependence relationship between *X* and *Y* and in particular the *Y*'s achievement of the task τ' depends on *X*. Even if it is important to analyze the nature of the *Y*'s delegation as for τ' [7] , in general *X* has the awareness to have any power on *Y* (and that *Y* believes this).
ii) if this delegation is spontaneous and in particular it is a special kind of delegation (for example it is a display of esteem) and *X* has awareness of this, i.e. *Bel (X* ($DoT_{YX\tau'} > \sigma_2$)) [8], in general an abstract benevolence could arise in *X* as for *Y*.
iii) *X* could imply from point (i) the *Y*'s unharmfullness (if the delegation nature permits it).
iv) trusting as a sign of goodwill and trustwortiness: Agents with bad intentions are frequently diffident towards the others; for example, in contracts they specify and check everything. Since they have non-benevolent intentions they ascribe similar attitudes to the others. This is why we believe that malicious agents are usually diffident and that (a risky abduction) suspicious agents are malicious. On such a basis, we also feel more trusting towards a non-diffident, trusting agent: this is a sign for us that it is goodwilling, non-malicious.

[6] where σ_1 is the *X*'s reasonable threshold for delegating to *Y* the task τ.

[7] for example, if there is already an agreement between *X* and *Y* about τ' with reciprocal commitments and possible sanctions in the case in which there could be some uncorrect behaviour.

[8] where σ_2 is the *Y*'s reasonable threshold for delegating to *X* the task τ'.

Table 6.4 How delegation can influence reciprocal trust

<table>
<tr><th></th><th>$DoT_{XY\tau} > \sigma_1$</th><th>$DoT_{XY\tau} < \sigma_1$</th></tr>
<tr><td rowspan="2">Bel(X S-Del(YXτ'))
and
Bel(X $DoT_{YX\tau'} > \sigma_2$)</td><td rowspan="2">Increases the X's degree of trust in Y as for τ</td><td>Increases the $DoT_{XY\tau}$ but
¬ Delegate (X Y τ)</td></tr>
<tr><td>Increases the $DoT_{XY\tau}$ and
Delegate (X Y τ)</td></tr>
<tr><td rowspan="2">Bel(X ¬S-Del(Y X τ'))
and
Bel(X $DoT_{YX\tau'} < \sigma_2$)</td><td>Decreases the $DoT_{XY\tau}$
but
Delegate (X Y τ)</td><td rowspan="2">Decreases the X's degree of trust in Y as for τ</td></tr>
<tr><td>Decreases the $DoT_{XY\tau}$
and
¬ Delegate (X Y τ)</td></tr>
</table>

Each of the previous points allows the possibility that *X* delegates to *Y* τ: Going to analyze the specific components of the degree of trust we can say that:

- the point (i) could increase both $DoW_{XY\tau}$ and $DoA_{XY\tau}$;
- the point (ii) and (iii) decrease the value of σ_1 and may be increase both $DoW_{XY\tau}$ and $DoA_{XY\tau}$.

In other words, after *Y*'s delegation we can have two new parameters:
$DoT'_{XY\tau}$ and σ_1' instaed of $DoT_{XY\tau}$ and σ_1 and it is possible that: $DoT'_{XY\tau} > \sigma_1$'. In Table 6.4 we display the different cases.

Another interesting case is when *Y*'s diffidence in *X* is believed by *X* itself:

Bel (X $DoT_{YX\tau'} < \sigma_2$*)* and for this

Bel (X ¬ *S-Delegates (Y X* τ'*))*

Also, in this case, various possible consequences in *X*'s mind are possible:

i) the fact that *Y* has decided not to depend on *X* could imply *Y*'s willingness to avoid possible retaliation by *X* itself; so that *X* could imply *Y* 's possible harmfulness.
ii) *Y* 's expression of lack of estimation of *X* and the fact that *Bel (X* $DoT_{YX\tau'} \leq \sigma_2$*)*, in general has the consequence that an abstract malevolence could arise in *X* as for *Y*.

Our message is trivial: the well-known interpersonal dynamics of trust (that trust creates a mirror trust, or diffidence induces diffidence), so important in conflict resolutions, in exchanges, in organizations, etc., must be grounded on their mental 'proximate mechanisms'; that is, on the dynamics of beliefs, evaluations, predictions, goals, and affects in *X* and *Y*' minds, and in *X*'s theory of *Y*'s mind and *Y*'s theory of *X*'s mind.

6.5 The Diffusion of Trust: Authority, Example, Contagion, Web of Trust

An interesting study about trust, partially different from the aspects so far analyzed, considers trust diffusion in groups and networks of agents as *meta-individual phenomenon*, and it is essentially based on problematics of contagion, for example, schemata, conventions, authority, and so on. In fact, trust is sometimes a *property of an environment*, rather than of a single agent or even a group: under certain conditions, the tendency to trust each other becomes diffused in a given context, more *like a sort of acquired habit or social convention than like a real decision*. These processes of 'trust spreading' are very powerful in achieving a high level of cooperation among a large population, and should be studied in their own right.

In particular, it is crucial to understand the subtle interaction between social pressures and individual factors in creating these 'trusting environments', and to analyze both advantages and dangers of such diffusive forms of trust.

In our analysis, these phenomena have also to be analyzed in terms of cognitive models, but in fact they follow slightly different rules from the ones considered in the individual interactions, and they concern a special kind and nature of goals and tasks.

In particular, we examine the point:

> *how widespread trust diffuses trust* (*trust atmosphere*), that is, how *X*'s trusting in *Y* can influence *Z* trusting *W* or *W'*, and so on. Usually this is a macro-level phenomenon and the individual agent does not calculate it.

Let us consider two prototypical cases, the two micro-constituents of the macro process:

i) Since *X* trusts *Y*, also *Z* trusts *Y*
ii) Since *X* trusts *Y*, (by analogy) *Z* trusts *W*.

We would like to underline the potential multiplicative effects of those mechanisms/rules: the process described in (i) would lead to a trust network like in Figure 6.9, while the process described in (ii) would lead to a structure like Figure 6.10.

6.5.1 Since Z *Trusts* Y*, Also* X *Trusts* Y

There are at least two mechanisms for this form of spreading of trust.

Agent's Authority (Pseudo-Transitivity)

This is the case in which, starting from the trustfulness that another agent (considered as an authority in a specific field) expresses about a trustee, other agents decide to trust that trustee.

Consider the situation:

$$Bel\ (X\ DoT_{ZY\tau} > \sigma_2)\ (where\ DoT_{ZY\tau}\ \text{is the degree of trust of}\ Z\ \text{on}\ Y\ \text{about the task}\ \tau)$$

that is:
agent *X* believes that *Z*'s degree of trust in *Y* on the task τ is greater than a reasonable threshold (following *Z*).

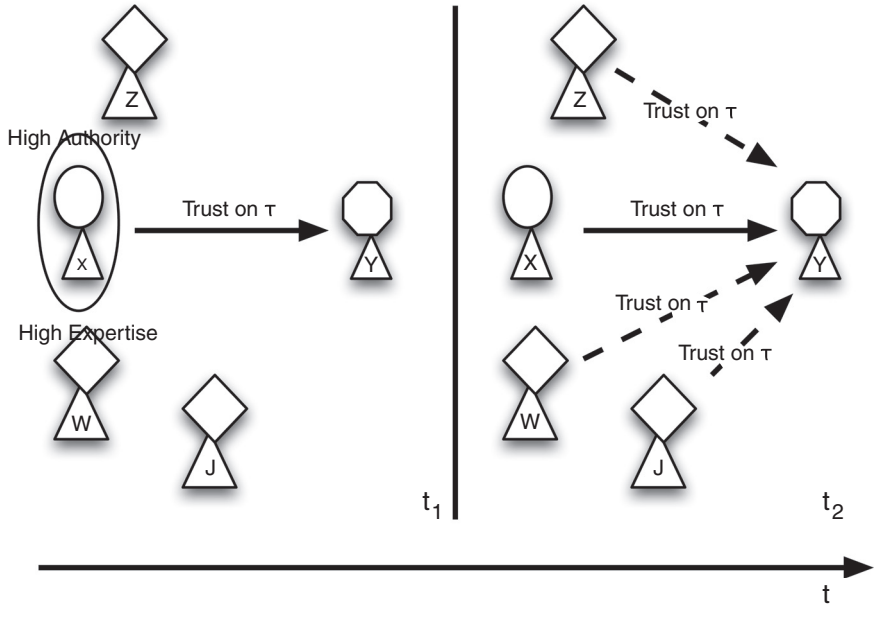

Figure 6.9 Emerging Trust in a specific trustee

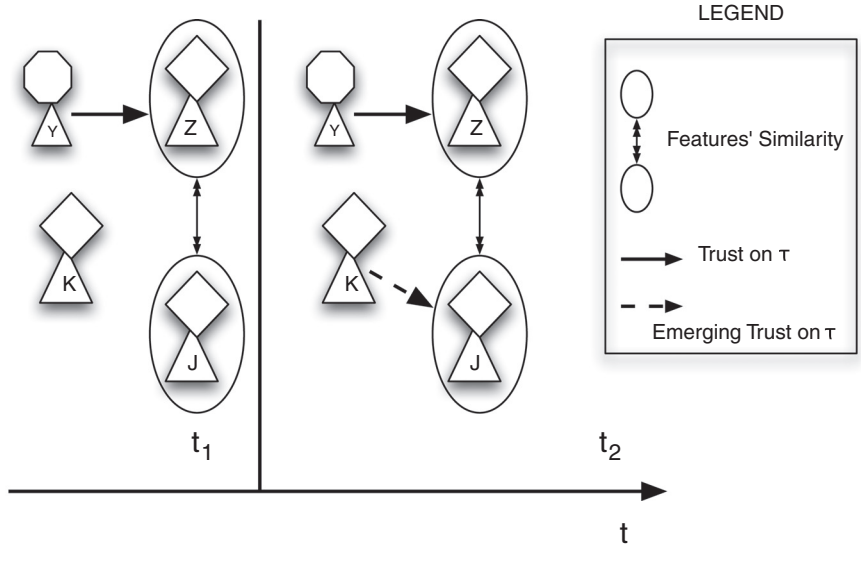

Figure 6.10 Emerging Trust by Analogy

Given this is X's belief, the question is: what about the $DoT_{XY\tau}$?

Is there a sort of *trust transitivity*? If so, what is its actual nature? (A frequent simplification about the trust transitivity is to consider it as a trivial question: if A trusts B and B trusts C, then A trusts C; in fact the problem is more complex and relates to the identification of the *real object to transfer* from an agent to another one).

Let us consider the case in which the only knowledge that X has about Y and about Y's possible performance on τ is given by Z. We hypothesize that:

$$DoT_{XY\tau} = Inf_X(Z\ Y\ ev(\tau)) * DoT_{XZev(t)}$$

where:

$DoT_{XZev(\tau)}$ is the X's degree of trust in Z about a new task $ev(\tau)$ that is the task of evaluating competences, opportunities, etc. about τ (see Figure 6.11). And $Inf_X(Z\ Y\ ev(\tau))$ represents

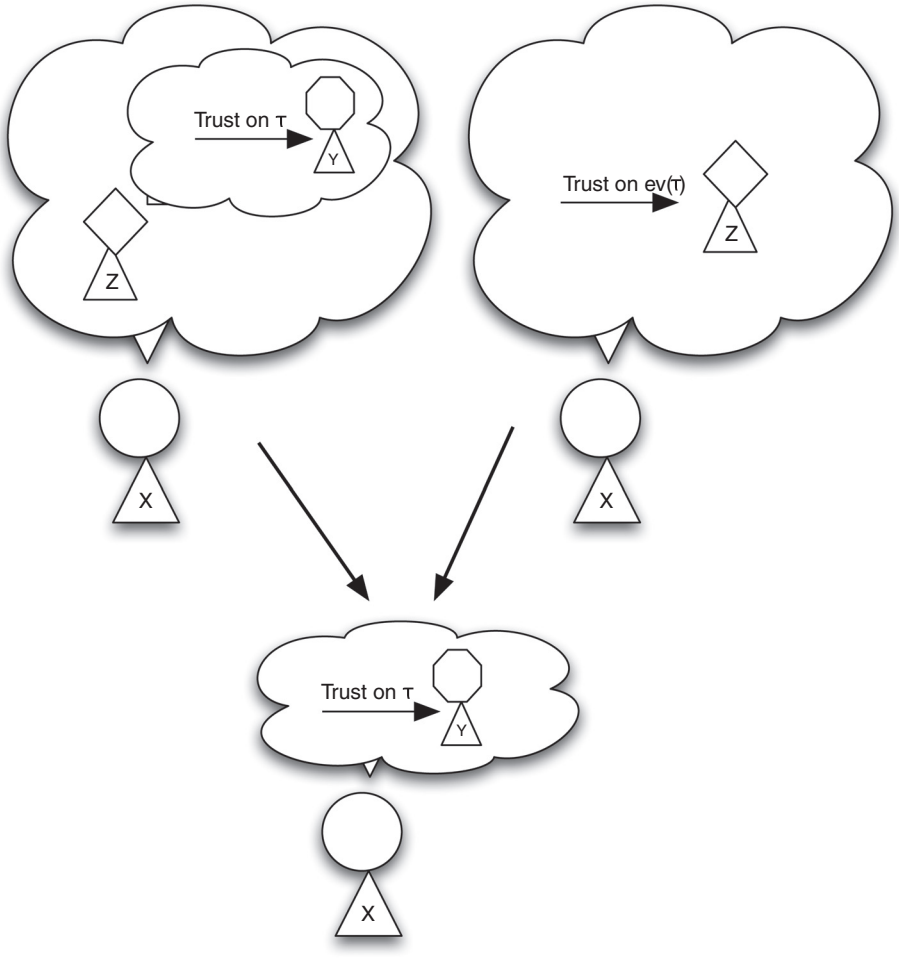

Figure 6.11 Pseudo-Transitivity: trusting through the others' trust

Z's potential influence in the task $ev(\tau)$ when τ is performed by Y (following X's opinion). This factor gives account of the different potential relationships among the (evaluating and evaluated) agents that can have an influence on the objectivity of the judgment of evaluation.

Notice that this pseudo-transitivity depends on subtle cognitive conditions. In fact, it is not enough that X trusts Z for adopting its trusting attitude towards Y; for us there is no real *transitivity* in trust: that X trusts Z and Z trusts Y does not imply that X trusts Y. Don't forget that trust is 'about' something; X's trust for Y is to do with some power, action, service. It has four arguments: not only X and Y but the task/goal and the context. About what does X trust Z? And about what does Z trust Y?

Suppose for example that X trusts Z as a medical doctor but considers Z a very impractical person as for business, and X knows that Z trusts Y as a stock-market broker agent; X has no reason to trust Y. On the contrary, if following X Z is a good doctor and he trusts Y as a good nurse, X can be learning to trust Y as a good nurse. What does this mean? This mean that X trusts Z as for a given *competence* in a given *domain* (Castelfranchi and Falcone, 1998b): Z is considered by X an 'authority' (or at least a good evaluator, expert) in this domain, this is why X considers Z's evaluation as highly reliable (Z is a trustworthy 'source' of evaluation).

So, since X trusts Y on the task τ we should have:

$$DoT_{XY\tau} = Inf_X(Z\ Y\ ev(\tau)) * DoT_{XZev(\tau)} > \sigma_1,$$

where σ_1 is the X's reasonable trust threshold for delegating τ.

We have a *pseudo-transitivity* (we consider it as cognitively-mediated transitivity) when:

1) $Inf_X(Z\,Y\,ev(\tau)) = 1$, in other words, there is no influence on Z by Y performing τ (following X) that could produce a non-objective judgment; and
2) $DoT_{XZev(\tau)} = DoT_{XZ\tau}$, in words, X has the same degree of trust in Z both on the task τ and on the task $ev(\tau)$; in other terms, X attributes the same DoT to Z both performing τ and trusting another agent about τ. This is a very common case in human activities, very often due to a superficial analysis of the trust phenomenon.

If the two above conditions are satisfied we can say that each time

$$DoT_{ZY\tau} > \sigma_2 \quad \text{and} \quad DoT_{XZ\tau} > \sigma_1 \text{ we will also have } DoT_{XY\tau} > \sigma_1.$$

In Table 6.5 we show the various possibilities of X's trusting or not of Y and how the two *DoTs* are combined.

Conformism

This mechanism is not based on a special expertise or authority of the 'models': they are not particularly expert, they must be numerous and just have experience and trust: since they do, I do; since they trust, I trust.

The greater the number of people that trust, the greater my feeling of safety and trust; (less perceived risk) the greater the perceived risk, the greater the necessary number of 'models' (Figure 6.12).

A good example of this is the use of credit cards for electronic payment or similar use of electronic money. It is a rather unsafe procedure, but it is not perceived as such, and this is

Table 6.5 Pseudo-Transitivity: Combining the Degrees of Trust

	$DoT_{ZY\tau} > \sigma_2$	$DoT_{ZY\tau} < \sigma_2$
$DoT_{X\,Z\,(ev)\tau} > \sigma_1$	X trusts Y as for τ	X does not trust Y as for τ
$DoT_{X\,Z\,(ev)\tau} < \sigma_1$	X trusts Y OR X does not trust Y as for τ	X trusts Y OR X does not trust Y as for τ

mainly due both to the utility and diffusion of the practice itself: it can be a simple conformity feeling and imitation; or it can be an explicit cognitive evaluation such as: 'since everybody does it, it should be quite safe (apparently they do not incur systematic damages)'. If everybody violates traffic rules, I feel encouraged to violate them too.

More formally, we have:

given $S=\{s_1,.\,.,\,s_n\}$ we are considering the set of agents belonging to the community as reference for trust diffusion.

if Bel $(X\,(DoT_{ZY\tau} > \sigma_2) \cap (DoT_{Z'Y\tau} > \sigma_3) \cap \ldots . \ldots \cap (DoT_{Z^nY\tau} > \sigma_n))$ then

$DoT_{XY\tau} > \sigma_1$.

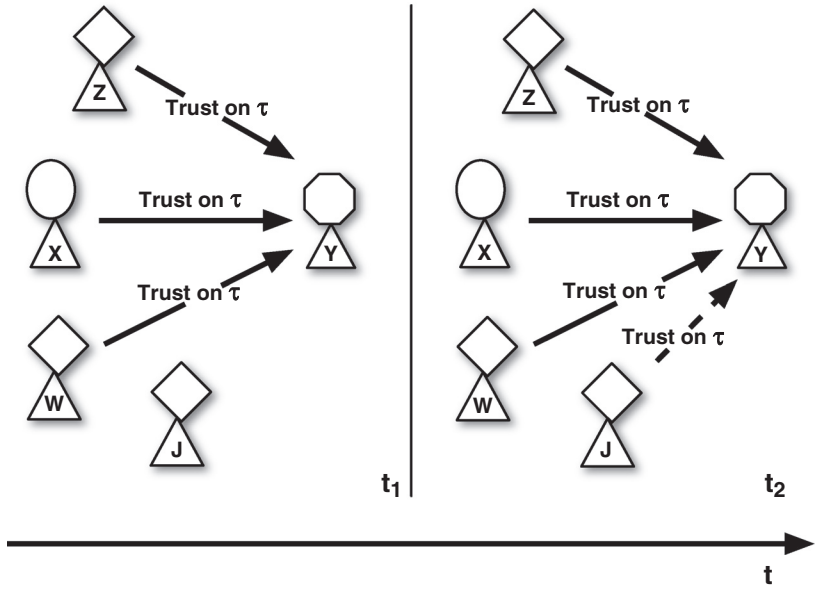

Figure 6.12 Trust Conformism

With $X, Z, . ., Z^n\ Y \in S$ *and* $\sigma_1, .., \sigma_n$ respectively their thresholds for delegating.

A very crucial example of this trust-spreading mechanism is about circulating information.

The greater the number of (independent) sources of a piece of information, a reputation, a belief, the more credible it is. (Castelfranchi, 1996)

In fact, belief acceptance is not an automatic process, it is subject to some tests and checks: one check is for its plausibility, coherence with previous knowledge, for the source reliability, etc. Thus, if several cognitive agents believe bel_1, probably bel_1 has been checked by each of them against its own direct experience, previous knowledge, and source evaluation, thus it is reasonable that it is more credible. Even in science, convergence is a criterion of validity. However, this is also a rather dangerous mechanism (of social bias and prejudice) since in fact sources in a social context are not independent and we cannot ascertain their independence, thus the number of sources can be just an illusion: there could be just a unique original source and a number of 'reporters'.

All these forms of trust spreading are converging in a given *target* of trust (being it an information source, a technology, a social practice, a company, a guy).

6.5.2 *Since* X *Trusts* Y, *(by Analogy)* Z *Trusts* W

This is a generalized form of trust contagion. It can be based on cognitive analogy or co-categorization:

- since *X* trusts *Y* (and *X* is expert) and *W* is like *Y*, *Z* will trust *W*
  or
- since everybody trusts some *Y*, and *W* is like/a *Y*, *Z* will trust *W*.

Where 'like' either means '*W* is similar to *Y* as for the relevant qualities/requirements', or means '*W* is of the same category of *Y*'; and 'some *Y*' means 'someone of the same category of *Y* or similar to *Y*'. Also in this case either a specific model and example counts, or a more generalized practice. For a more diffused and analytic treatment of this argument see the following Section 6.6.

6.5.3 *Calculated Influence*

As we said, usually, this trust contagion is a macro-level phenomenon and the individual agents do not calculate it, it is not intended. However, sometimes it is a strategic decision. Suppose for example that *X* believes herself to be a 'model' for *Z*, and that she wants *Z* to trust *Y* (or *W*). In this case *X* can deliberately make *Z* believe that he trusts *Y*, in order to influence *Z* and induce him to trust *Y* (or *W*).

The smart businessman might be aware of the fact that when he buys certain shares, a lot of other people will follow him, and he can exploit precisely this imitative behavior of his followers and speculate at their expense.

Trust Atmosphere

All the previous mechanisms (Section 6.5.1 through to Section 6.5.3) are responsible for that celebrated *'trust atmosphere'* that is claimed to be the basis of a growing market economy or of a working political regime. They are also very fundamental in computer mediated organizations, interactions (like electronic commerce), cooperation (CSCW), etc. and even in multi-agent systems with autonomous agents.

The emergence of a *'trust atmosphere'* just requires some generalization (see later Section 6.6): the idea (or feeling) not simply that 'X trusts Y' or/and 'Z trusts X/Y', but that 'everybody (in this context) trusts Y' or 'Z trusts everybody', or even 'here everybody trusts everybody'; or at least that 'a lot of people trust Z', 'a lot of people trust a lot of people here'.

6.6 Trust Through Transfer and Generalization

Would you buy a 'Volkswagen Tomato sauce'? And what about a FIAT tomato sauce? Why is it a bit more credible? And what about a Volkswagen water pump? Why is it appealing, perceived as reliable? How marketing managers and advertising people decide to sell a new product (good, service) under a well known brand? This is a serious problem in marketing, but, not based on a real model.

Would you fully trust a surgeon to recommend you a medication? Or a simple medical doctor to perform a serious surgical intervention? In general: *If X trusts Y for a given τ, will she Trust Y also for τ'?*

In this section, we analyze how it is possible to predict how/when an agent who trusts something/someone will therefore trust something/someone else, before and without direct experience. This is different from the models of trust just based on (or reduced to) a probability index or a simple measure of experience and frequency, we are interested in analyzing the trust concept so that we are able to cope with problems like: given X's evaluation about Y's trustworthiness on a specific task τ, what can we say about X's evaluation of Y's trustworthiness on a different but analogous task τ'? Given X's evaluation of Y's trustworthiness on a specific task τ, what can we say about X's evaluation of the trustworthiness of a different agent Z on the same task τ?

In fact, in our view only a cognitive model of trust, with its analytical power (as showed in Chapter 2), seems able to account for the inferential generalization of trustworthiness from task to task and from agent to agent not just based on specific experience and/or learning.

In general, trust derives, directly or indirectly, from the experience.[9] There are computational models of trust in which trust is conceived as an expectation sustained by the repeated direct interactions with other agents under the assumption that iterated experiences of success strengthen the trustor's confidence (Witkowski *et al.*, 2001), (Jonker and Treur, 1999). In the case of indirect experience, the more diffused case of study is the trust building on the basis of the others' valuations (reputation) ((Sabater and Sierra, 2001), (Jurca and Faltings, 2003), (Conte and Paolucci, 2002). A different and also interesting case of indirect experience for trust building (in some cases we can speak of attempts to rebuild, by other tools than observability, the direct experience), not particularly studied in these years, is based on the

[9] We have to say that there is also a part of trust that derives from some personality-based or cultural 'disposition' not based on previous experience.

inferential reasoning over the categories on which the world is organized (or could be thought to be organized): real and mental categories.

6.6.1 Classes of Tasks and Classes of Agents

In our model of trust we consider the trustor (X) and the trustee (Y) as single agents,[10] and the task (τ) as a specific task. For reasons of generality, optimization, economy, and scalability it would be useful to apply the trust concept not only to specific tasks and to single agents. In fact, it would be really useful and realistic to have a trust model that permits trust to be transferred among similar agents or among similar tasks. In this sense having as reference classes of tasks and classes of agents (as humans generally have) would be extremely important and effective. A good theory of trust should be able to understand and possibly to predict how/when an agent who trusts something/someone will therefore trust something/someone else, and before and without a direct experience. And, vice versa, from a negative experience of trustworthiness it could be possible to extract elements for generalizing about tasks and/or agents.

In this perspective we have to cope with a set of problems (grouped in two main categories):

1) Given X's evaluation about Y's trustworthiness on a specific task τ, what can we say about X's evaluation of Y's trustworthiness on a different but analogous task τ'? What should we intend for an 'analogous task'? When does the analogy work and when does it not work between τ and τ'? How is it possible to modify X's evaluation about Y's trustworthiness on the basis of the characteristics of the new task? How can we group tasks in a class? And so on.
2) Given X's evaluation about Y's trustworthiness on a specific task (or class of tasks) τ, what can we say about X's evaluation of the trustworthiness of a different agent Z on the same task (or class of tasks) τ? Which are the agent's characteristics that transfer (or not) the evaluation to different trustees?

In fact, these two sets of problems are strictly interwined with each other and their solutions require a more careful analysis of the nature of tasks and agents.

6.6.2 Matching Agents' Features and Tasks' Properties

In general, we can say that if an agent is trustworthy with respect to a specific task (or class of tasks) it means that, from the trustor's point of view, the agent has a set of *specific features* (resources, abilities and willingness) that are useful for that task (or class of tasks). But, what does it mean: useful for that task? We can say that, again depending on the trustor's point of view, a task has a set of *characterizing properties* requiring specific resources and abilities of various natures, which can be matched in some way with the agents' features. The attribution of the features to the agents, the right individuation of the tasks' properties and the match between the first and the second ones represent different steps for the trust building and are the bases for the most general inferential reasoning process for the trust generalization phenomenon.

[10] Either an 'individual' or a 'group' or an 'organization'.

The above described three attributions (features, properties and match) are essential for the success of trust building. For example, imagine the task of 'taking care of a baby during evening' (trustor: *baby's mother*; trustee: *baby-sitter*).

The *main properties* of the task might be considered:

a) to avoid dangers to the children;
b) to satisfy childrens' main physical needs;
c) to maintain a peaceful and reassuring climate by playing.

At the same time, we could appreciate several *main features* of the trustee:

1) careful and scrupulous;
2) lover of children;
3) able to maintain concentration for long time;
4) proactive;
5) impulsive, agitated and nervous.

The operation for evaluating the adequacy of the trustee to the task is mainly based on the match between the trustee features (that become '*qualities*' or '*defects*'; see Chapter 2) and the properties of the task. In the example, we can say that the feature number *(1)* is *good* for satisfying the properties *(a)* and *(b)*; the feature number *(2)* is *good* for satisfying the properties *(b)* and *(c)*; the feature numbers *(3 and 4)* are *good* for satisfying the properties *(a)* and *(b)*; the feature number *(5)* is *bad* for satisfying the properties *(a)* and *(c)*.

Both the properties of the task and the features of the trustee could be perceived from different trustors in different ways (think about the possible discussions in real life between a mother and a father about this). Not only this: the match could also be considered in a different way from different personalities and point of views. In addition, both the features of an agent and the properties of a task can be considered unchanged or not during time: it depends on the tasks, on the trustees and the trustors' perception/representation.

It is superfluous to be reminded that this kind of trust building is just one of the many ways in which to definine the agents' trustworthiness. Sometimes, the trustors do not know, except at a superficial level, the tasks' properties and/or the trustees' features (like when trust building is based on reputation or many cases of direct experiences).

The trust building based on the main inferential reasoning process, is then depending on several different factors and on their composition. When inferring the task's properties a trustor has to select the minimal acceptable values for the included indispensable ones (if there are any). At the same time, the trustor has to evaluate the potential trustee's features and verify their compatibility and satisfaction for the given task. These are complex attributions depending on the trustor and her trust model.

Starting from this kind of attribution we will analyze the phenomenon of generalization between similar tasks and similar agents. The essential informal 'reasoning' one should model can be simplified as follows:

- To what features/qualities of Y (the trustee) is its validity ascribable for the requirements of τ?

- Has Z the same relevant qualities? How much; how many? Does Z belong to the same class/set of Y, based on the relevant features?
- Does τ' share the same relevant requirements as τ? Does τ' belong to the same kin/class/set of services/goods as τ ?

6.6.3 Formal Analysis

In more systematic and formal terms we have tasks (τ) and agents (Ag): with $\tau \in T \equiv \{\tau_1, \ldots, \tau_n\}$, and $Ag \in AG \equiv \{Ag_1, \ldots, Ag_m\}$. We can say that each task τ can be considered composed of both a set of actions and the modalities of their running, this we call *properties*:

$$\tau \equiv \{p_1, \ldots, p_n\} \tag{6.14}$$

we consider this composition from the point of view of an agent (Ag_X):

$$Bel_{AgX}\ (\tau \equiv \{p_1, \ldots, p_n\}) \tag{6.15}$$

In general each of these properties could be evaluated with a value ranging between a minimum and maximum (i.e.: *0,1*): representing the complete failure or the full satisfaction of that action. So in general: $0 \leq p_i \leq 1$ with $i \in \{1, .. , n\}$.

Of course, not all the properties of a task have the same relevance: some of them could be considered indispensable for the realization of the task, others could be considered useful in achieving greater success.

If we insert an apex c to all the properties that the agent (Ag_X) considers indispensable (*core properties*) for that task, we can write:

$$Bel_{AgX}\ (\tau \equiv \{p_1^c, \ldots, p_k^c\} \cup \{p_1, \ldots, p_m\}) \tag{6.16}$$

We call $\tau^C \equiv \{p^c{}_1, \ldots, p^c{}_k\}$ and $\tau^{NC} \equiv \{p_1, \ldots, p_m\}$, so

$$\tau = \tau^C \cup \tau^{NC} \tag{6.17}$$

The set of the core properties is particularly relevant for grouping tasks into classes and for extending the reasoning behind generalization or specification.

Analogously, we can define the set of the *features* f_{AgY} for an agent (Ag_Y):

$$f_{AgY} \equiv \{f_{Y1}, \ldots, f_{Yn}\} \tag{6.18}$$

we consider this composition from the point of view of an agent (Ag_X):

$$Bel_{AgX}(f_{AgY} \equiv \{f_{Y1}, \ldots, f_{Yn}\}) \tag{6.19}$$

Also in this case, each of these features could be evaluated with a value ranging between a minimum and a maximum (i.e.: *0,1*): representing the complete absence or the full presence of that feature in the agent Ag_Y, from the point of view of Ag_X. So in general: $0 \leq f_i \leq 1$ *with* $i \in \{Y_1, \ldots, Y_n\}$.

Given the previous definitions, we can say that Ag_X could trust Ag_Y on the task τ if Ag_X has the following beliefs: (6.19), (6.16) and

$$Bel_{AgX}\ (\forall p_i \in \tau^C,\ \exists\{f_i\} \in \{f_{Y1}, \ldots, f_{Yn}\} | \{f_i\} \neq \emptyset \cap p_i\ \textit{is satisfied from}\ \{f_i\}) \tag{6.20}$$

where 'p_i *is satisfied from* $\{f_i\}$' means that the trustee (from the point of view of the trustor) has the right features for satisfying all the core properties of the task. In particular, the needed features are over the minimal threshold (σ_j) so that the main properties of the task will also be over the minimal threshold (ρ_i):

$$(\forall f_j \in \{f_i\},\ (f_j > \sigma_j))\ and\ (apply\ (\{f_i\},\ \tau) \Rightarrow (\forall p_i \in \tau^C,\ p_i > \rho_i)) \tag{6.21}$$

where the *apply* function defines the match between the agent's features and the task's properties. Also the different thresholds (σ_j and ρ_i) are depending on the trustor.

We have to say that even if it is possible to establish an objective and general point of view about both the actual composition of the tasks (the set of their properties, including the actual set of core properties for each task) and the actual features of the agents, *what is really important are the specific beliefs of the trustors about these elements*. In fact, on the basis of these *beliefs* the trustor would determine its trust. For this reason alone we have introduced as main functions those regarding the trustors' beliefs.

6.6.4 Generalizing to Different Tasks and Agents

Let us now introduce the reasoning-based trust generalization. Consider three agents: Ag_X, Ag_Y and Ag_Z (all included in AG) and two tasks τ and τ' (both included in T). Ag_X is a trustor and Ag_Y and Ag_Z are potential trustees.

Where:

$$\tau' \equiv \{p'^c_1, \ldots,\ p'^c_k\} \cup \{p'_1, \ldots,\ p'_m\} = \tau'^C \cup \tau'^{NC}) \tag{6.17bis}$$

and in general

$$(p'_j \neq p_j)\ with\ p'_j \in (\tau'^C \cup \tau'^{NC})\ and\ p_j \in (\tau^C \cup \tau^{NC});$$

$$Ag_Z \equiv f_{AgZ} \equiv \{f_{Z1}, \ldots, f_{Zn}\} \tag{6.18bis}$$

The first case (*caseA*) we consider is when Ag_X does not know either the τ's properties, or the Ag_Y features, but they trust Ag_Y on τ (this can happen for different reasons: for example, he was informed by others about this Ag_Y's trustworthiness, or simply he knows the successful result without assisting in the whole execution of the task, and so on). In more formal terms:

a1) $Trust_{AgX}(Ag_Y, \tau)$
a2) $\neg Bel_{AgX}(f_{AgY} \equiv \{f_{Y1}, \ldots, f_{Yn}\})$
a3) $\neg Bel_{AgX}(\tau \equiv \{p^c_1, \ldots, p^c_k\} \cup \{p_1, \ldots, p_m\})$

In this case which kind of trust generalization is possible? Can Ag_X believe that Ag_Y is trustworthy on a different (but in some way analogous) task τ' (generalization of the task) starting from the previous cognitive elements (*a1, a2, a3*)? Or, can Ag_X believe that, another different (but in some way analogous) agent Ag_Z is trustworthy on the same task τ (generalization of the agent), again starting from the previous cognitive elements (*a1, a2, a3*)?

The problem is that the *analogies* (between τ and τ', and between Ag_Y and Ag_Z) are not available to Ag_X because they do not know either the properties of τ or the features

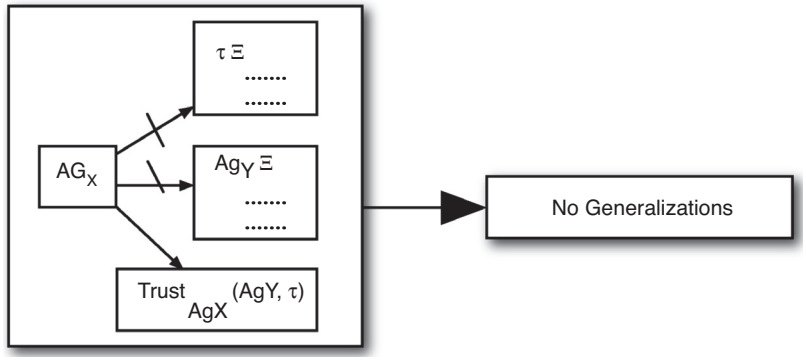

Figure 6.13 Generalization in case of Ag_X's ignorance about task's properties and trustee's features. (Reproduced with kind permission of Springer Science+Business Media © 2008)

of Ag_Y. So we can conclude (see Figure 6.13) that in *caseA* there is no possible rationale, grounded generalization to other tasks or agents: in fact the set of *(a1, a2, a3) do not permit generalizations.*

In fact, the only possibility for a rational generalization in this case is given from an indirect generalization by Ag_X . For example someone else can suggest to Ag_X that on the basis of his first trustworthy attitude he can trust another agent or another task because there is an analogy between them: in this case Ag_X, trusting this suggestion, acquires the belief for an indirect generalization. The second case (*caseB*) we consider is when Ag_X does not know Ag_Y's features, but he knows τ's properties, and he trust Ag_Y on τ (also in this case for several possible reasons different from inferential reasoning on the match between properties and the features). In more formal terms:

b1) $Trust_{AgX}(Ag_Y, \tau)$
b2) $\neg Bel_{AgX}(f_{AgY} \equiv \{f_1, \ldots, f_n\})$
b3) $Bel_{AgX}(\tau \equiv \{p_1^c, \ldots, p_k^c\} \cup \{p_1, \ldots, p_m\})$

Despite the ignorance about Ag_Y's features (*b2*), Ag_X can believe that Ag_Y is trustworthy on a different (but in some way analogous) task τ' (generalization of the task) just starting from the previous cognitive elements (*b1* and *b3*) and from the knowledge of τ''s properties. He can evaluate the overlap among the core properties of τ and τ' and decide if and when to trust Ag_Y on a different task. It is not possible to generalize with respect to the agents because there is no way of evaluating any analogies with other agents.

So we can conclude (see Figure 6.14) that in the case (*B*) task generalization is possible: in fact the set *(b1, b2, b3) permits task generalizations but does not permit agent generalizations.*

Also, in this case we can imagine an *indirect* agent generalization. If Ag_X trusts a set of agents $AG1 \equiv \{Ag_Y, Ag_W \ldots, Ag_Z\}$ on a set of different but similar tasks $T1 \equiv \{\tau, \tau', \ldots, \tau^n\}$ (he can evaluate this similarity given his knowledge of their properties) he can trust each of the agents included in *AG1* on each of the tasks included in *T1*.

The third case (*caseC*) we consider is when Ag_X does not know τ's properties, but he knows Ag_Y's features, and he trusts Ag_Y on τ (again for several possible reasons different from

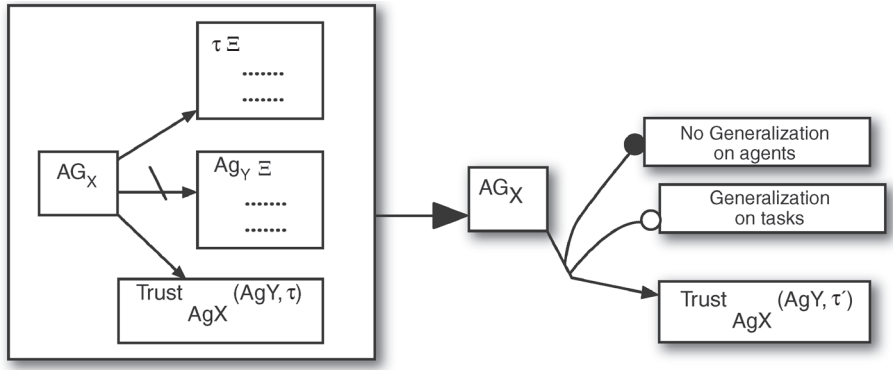

Figure 6.14 Generalization in case Ag_X knows only the trustee's features. (Reproduced with kind permission of Springer Science+Business Media © 2008)

inferential reasoning on the match between properties and the features). In more formal terms:

c1) $Trust_{AgX}(Ag_Y, \tau)$
c2) $Bel_{AgX}(f_{AgY} \equiv \{f_1, \ldots, f_n\})$
c3) $\neg Bel_{AgX}(\tau \equiv \{p_1^c, \ldots, p_k^c\} \cup \{p_1, \ldots, p_m\})$

Despite the ignorance about τ's features (*c3*), Ag_X can believe that a different (but in some way analogous) agent Ag_Z is trustworthy on the task τ (generalization of the agent) just starting from the previous cognitive elements (*c1* and *c2*) and from the knowledge of Ag_Z's features. He can evaluate the overlap among the features of Ag_Y and Ag_Z and decide if and when to trust Ag_Z on τ. While, in this case, it is not possible to generalize a task because there is no way of evaluating any analogies with other tasks.

So we can conclude (Figure 6.15) that in the case (*C*) agent generalization is possible: in fact the set *(c1, c2, c3) permits agent generalizations but does not permit task generalizations*.

Exactly as in case *B*, also in this case we could imagine an *indirect* task generalization. If Ag_X trusts a set of different but similar agents $AG1 \equiv \{Ag_Y, Ag_W \ldots, Ag_Z\}$ (he can evaluate this

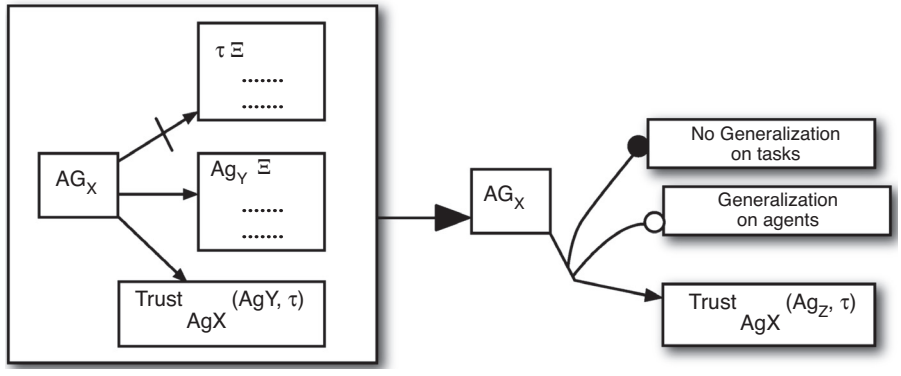

Figure 6.15 Generalization in case Ag_X knows only the task's properties. (Reproduced with kind permission of Springer Science+Business Media © 2008)

similarity given his knowledge of their features) on a set of tasks $T1 \equiv \{\tau, \tau', \ldots, \tau^n\}$ he can trust each of the agents included in *AG1* on each of the tasks included in *T1*.

The fourth case (*caseD*) we consider is when Ag_X knows both τ's properties and Ag_Y's features, and they trust Ag_Y on τ (in this case for inferential reasoning on the match between properties and the features). In more formal terms:

d1) $Trust_{AgX}(Ag_Y, \tau)$
d2) $Bel_{AgX}(f_{AgY} \equiv \{f_1, \ldots, f_n\})$
d3) $Bel_{AgX}(\tau \equiv \{p_1^c, \ldots, p_k^c\} \cup \{p_1, \ldots, p_m\})$

Ag_X can both believe:

- that a different (but in some way analogous) agent Ag_Z is trustworthy on the task τ (generalization of the agent) starting from the cognitive elements (d1 and d2) and from the knowledge of Ag_Z's features; or
- that Ag_Y is trustworthy on a different (but in some way analogous) task τ' (generalization of the task) starting from the cognitive elements (d1 and d3) and from the knowledge of τ' properties; or again
- that a different (but in some way analogous) agent Ag_Z is trustworthy on a different (but in some way analogous) task τ' (generalization of both the agent and the task).

So we can conclude (see Figure 6.16) that in the case (*D*) both task and agent generalization is possible: in fact the set *(d1, d2, d3) permits both agent and task generalizations.*

Note that we have considered in all the studied cases (*a1, b1, c1*, and *d1*) the fact that Ag_X *trusts* Ag_Y on the initial task τ. In fact, in case of *distrust* (or mistrust) we could receive analogous utility (in cases *B*, *C* and *D*) for the distrust (or mistrust) generalization to the other agents or tasks. In other words, Ag_X on the basis of his experience with Ag_Y on task τ could distrust agents similar to Ag_Y or/and tasks similar to τ.

The case in which, from a specific knowledge about agents, tasks and their matches, a trustor can infer complementary features or properties *contradicting* their previous positive or

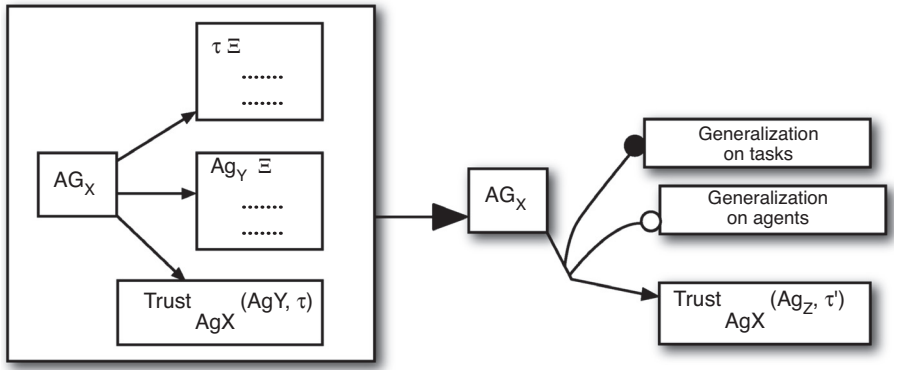

Figure 6.16 Generalization in case Ag_X knows both task's properties and trustee's features. (Reproduced with kind permission of Springer Science+Business Media © 2008)

negative beliefs is more complex. For this it is necessary a more subtle (fine-grained) analysis of features and properties.

Let us now try to refine the previous analysis. Starting from our characterization of the tasks, we can say that the similarity between different tasks is given by the overlap among the properties, in particular among their core properties (the indispensable relevant ones). So, the interesting thing is that, if the trustor believes that the trustee has the right features for a specific task, and another task has a significant overlap of its properties with that task (and specifically they share the core properties), then the trustor could also believe that the trustee has the right features for realizing the second task.

Also, in general, we can say that, given two tasks τ and τ' where both the formulas (6.17) and (6.17bis) are true, if the trustee (on the point of view of the trustor) has the right features for realizing τ and:

$$\tau'^{C} \subseteq \tau^{C} \tag{6.22}$$

then, on the trustor's point of view, that trustee also has the right features for realizing the task τ'.

So, even if the trustor had never seen the trustee operate on the task τ', it can infer its trustworthiness on that task. Note that the trustor could trust the trustee on a new task τ' also in the case in which they ignoring their features, but has experienced their trustworthiness about τ (in fact deducing them from the previous task in an indirect way).

In the case in which:

$$\tau'^{C} \not\subset \tau^{C} \tag{6.23}$$

there is at least one $p'_j \in (\tau'^{C})$ that is different from all the $p_j \notin (\tau^{C})$. It means that in this case the trustor can trust the trustee if: for each of the $p'_j \in (\tau'^{C})$ either it is equal to one of the $p_j \in (\tau^{C})$, or the trustor believes that it is satisfied from at least one of the trustee features ($f_{AgY} \equiv \{f_1, \ldots, f_n\}$). In other words, the trustor can trust the Ag_Y on a different task τ' if they believe that:

- either all the core properties of τ' are included in the core properties of τ (where Ag_Y is believed trustworthy on the task τ);
- or for each property of τ' not included in τ Ag_X believes there is at least one feature of Ag_Y that satisfies that property (and, of course, Ag_Y is believed to be trustworthy on the task τ).

Of course, all the considerations made in this paragraph can also be applied to self-confidence and self-efficacy; Ag_X's trust about himself: 'Given Ag_X's success on task τ will Ag_X be able on task τ' ?'; 'Given Ag_Y's success on τ will Ag_X be able on τ' too?'.

6.6.5 *Classes of Agents and Tasks*

It is quite simple to extend the above reasoning to a class of tasks. We can define a class of tasks as the set of tasks sharing a precise set of (quite abstract) core properties. The different specifications of these properties define the different tasks belonging to the class. On the basis of the core properties of a class, we can select all the agents endowed with the appropriate features who will be trustworthy for the whole class of tasks.

If we call *KT* a class of tasks, we can say that it is characterized from a set of (quite abstract) properties. The tasks belonging to class *KT* are all those whose core properties are specifications of the abstract properties of *KT*. In more formal terms, in addition to (6.17) we have:

$$KT \equiv \{ap_1, \ldots, ap_n\} \tag{6.24}$$

τ is a task belonging to the class *KT* if and only if:

$$(\forall ap_i \ (with\ n \geq i \geq 1)\ \exists p_j \ (with\ k \geq j \geq 1) | (p_j = ap_i) \vee (p_j\ is\ a\ \text{specification}\ of\ ap_i)) \tag{6.25}$$

where ap_i (*with n≥i≥1*) is an abstract property.

For example, we can say that the class of tasks of 'taking care of people' is constituted from the following abstract properties:

a) to avoid dangers to the people;
b) to satisfy peoples' main physical needs;
c) to satisfy peoples' main psychological needs.

In the case of the task 'taking care of children', we can say that it is included in the class of 'taking care of people' because all of its core properties are specializations of the abstract properties of the main task class:

a) to avoid dangers to children;
b) to satisfy childrens' main physical needs;
c) to maintain a peaceful and reassuring climate by playing.

At the same time we could consider 'to take care of elderly people' another task included in this class, because the core properties in this case are also:

a) to avoid dangers to the elderly;
b) to satisfy the elderlies' main physical needs;
c) to maintain a peaceful and reassuring climate;

are specifications of the same abstract properties of the class.

The situation is more complex when a task is just partially included in a class. Consider for example the case of 'to take care of a house'. The core properties are:

a) to clean and order the things in the house;
b) to avoid damage to the things.

In these cases the analysis has to consider the potential trustee's features and the classes of attitudes they have. The match among classes of agents and classes of tasks can inform us about positive or negative attitudes of the agents belonging to that agent class with respect to the tasks belonging to that task class. Of course, going towards specifications of both agent classes and task classes permits us to establish better matches and confrontations about which agent is more adequate for which task.

6.7 The Relativity of Trust: Reasons for Trust Crisis

As we saw so far in this chapter, trust dynamics is a complex and quite interesting phenomenon. Its diverse nature depends on the many and interacting elements on which trust is essentially based (as shown in this book). One of the main consequences of a complex trust dynamics is the fact that it produces what we generally call *crisis of trust*. And they could be different and articulated on the basis of the causes from which they derive.

To analyze the situations in which trust relationship can enter into a crisis (and in case to collapse) in depth we have to take into consideration the different elements we have introduced into our trust model as showed in Chapters 2 and 3. In fact, there are interesting and complex dynamical interactions among the different basic elements for trusting that have to be considered when evaluating the trust crisis phenomena.

We show the different trust crises starting from the basic trust model and increase it with the complexities needed to describe the complete phenomenon of trust.

As we saw, we call the trustor (X) with a goal (g_X) that she wants achieve by a delegation to another agent (Y) assigning to him the task (τ). This delegation is based on two of X's main beliefs:

i) The fact that Y has the features for achieving the task (in our model they are represented by competences, skills, resources, tools, but also by willingness, persistence, honesty, morality, and so on); these features must be sufficient to achieve the involved task (we made use of thresholds for measuring this sufficiency (see Chapter 3); and these thresholds were dependant on both the goal's relevance, the potential damages in case of failure, the personality of the agent, etc.).
ii) The fact that the task τ is believed appropriate for achieving a world state favouring (or directly achieving) the goal g_X (see Figure 6.17).

In this first simplified trust model, X can 'revise' her own trust in Y on τ (and then go in a trust crisis) on the basis of different reasons:

i) she can change her own beliefs on Y about his features (for example, X does not evaluate Y sufficiently able and/or motivated for the task τ);
ii) X can change her own beliefs about the appropriateness of the task τ for the achievement of the considered goal: maybe the action α is not useful (in X's view) for achieving the world state p; or may be p no longer satisfies the goal g_X;
iii) in X's mind the value (and then the relevance) of the goal g_X can change or it is suddenly achieved by other means.

In fact, a more developed and appropriate model of trust gives us additional elements for the analysis. We have to consider the constitutive components of the two main attitudes of Y (competence and willingness). As shown in Figure 2.15 in Chapter 2, there are many sources for these main beliefs. And they can change. At the same time, X could/should consider how the context/environment in which the task has to be realized changes in its turn: in this way introducing facilitant or impeding elements with respect to the original standard (or previously evaluated) situation (see Figure 6.18).

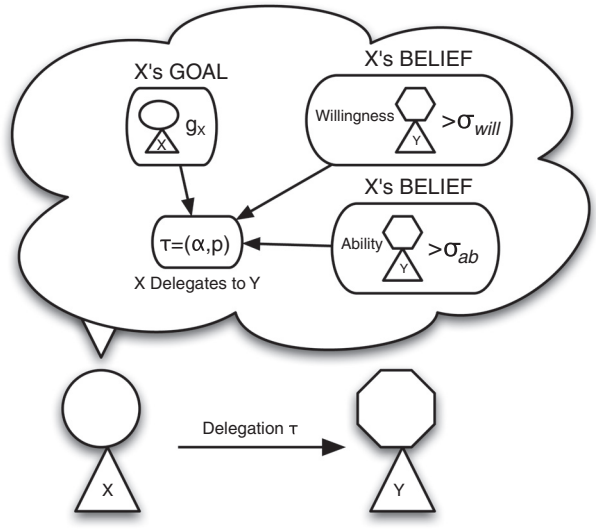

Figure 6.17 *X*'s Mental Ingredients for Delegating to *Y* the task τ

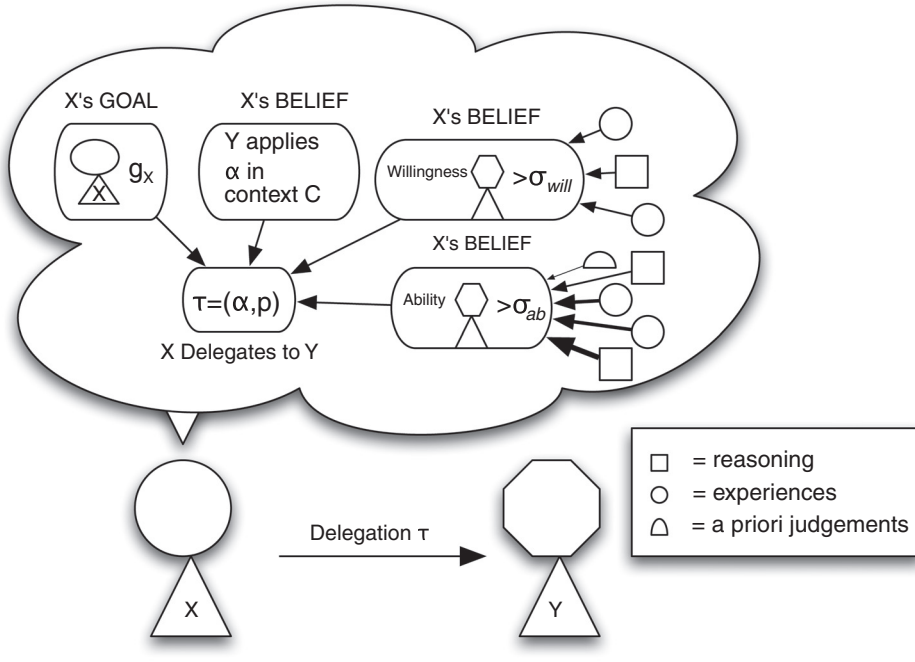

Figure 6.18 *X*'s Mental Ingredients for Delegating to *Y* the task τ (introducing more beliefs and the Context)

In this new picture (see Figure 6.18) *X*'s crisis of trust can be based on different causes:

i) due to a deeper analysis of *her own* beliefs about the reasons supporting *Y*'s willingness and abilities, *X* can evaluate (analyzing better strengths and weakness of her beliefs) the appropiateness of her previous judgment;
ii) due to a deeper analysis of *her own* beliefs about the appropiateness of the delegated task to the achievement of the goal; in particular in the better analyzed and defined context/environment (*X* realizes that (given that context) action α does not achieve the state *p*; and/or *p* does not include/determine g_X).
iii) due to the decrease of the goal's value g_X. In particular, the cost of delegation (in its general terms, not only in economic sense) is not comparable with (is not paid from) the achievement of the goal.

Again increasing and sophisticating the trust model we have the situation shown in Figure 6.19.

In this new scenario it is not only the presence of *Y* (with his features deeply analyzed) that is considered in the specific contex/environment, but also the availabilty of other potential trustees (*Z*, *W* in Figure 6.19). Their presence represents an additional opportunity for *X* (depending from *X*'s judgments about *W* and *Z* and about their potential performances in the specific environment) of achieving her own goal. In fact this opportunity can elicit *X*'s crisis

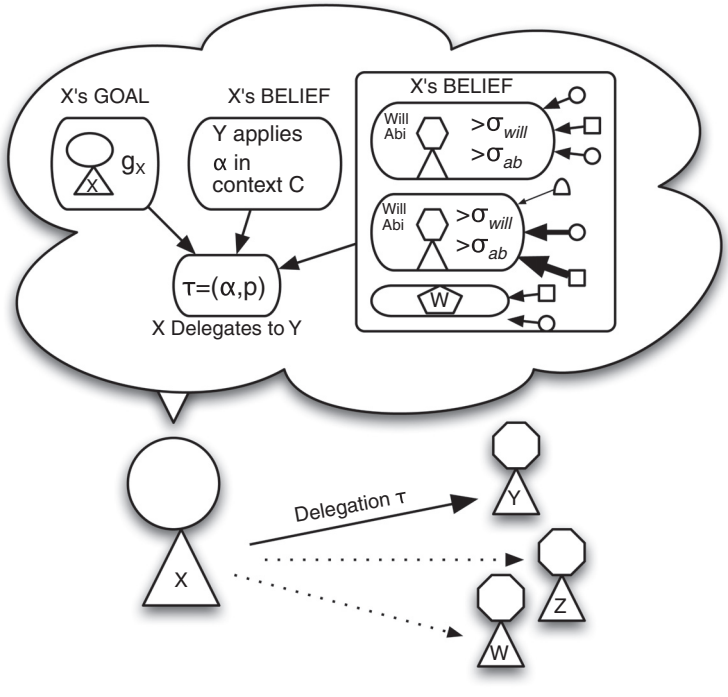

Figure 6.19 *X*'s Mental Ingredients for Delegating to *Y* the task τ (considering also other potential trustees)

of trust in Y. This kind of crisis is a bit different from those previously shown: X could again evaluate Y as able to realize the task, but she considers whether other available agents would be more suitable delegates (W or Z).

We can resume the reasons of this crisis in the following way:

i) given new X's beliefs on Y's features, on the context, and on features of other agents, we can have a trust crisis in Y with delegation of the task to another agent (Z or W in Figure 6.19).
ii) due to a deeper analysis of *her own* beliefs about the appropiateness of the delegated task to the achievement of the goal; in particular in the better analyzed and defined context/environment (X realizes that (given that context) action α does not achieve the state p; and/or p does not include/determine g_X). These considerations are also true in the case in which there is more than one agent available for delegation: in fact, there is no added value (nor is there unsufficient added value) with the presence of more potential trustees.
iii) due to the decrease of the goal's value g_X. In particular, the cost of delegation (in its general terms, not only in economic sense) is not comparable with (is not paid from) the achievement of the goal. And this is true even in the case of a presence of more potential trustees.

In the final, more complete, version of the trust model we introduce the question of the contemporary presence of X's goals, evaluating the competition among them and the consequent dynamics for the evaluation of the priorities (Figure 6.20).

In this case the decrease of a goal's value (say g) and the increase of the value of another one (say g') could elicit X's trust crisis toward Y with respect to τ. This crisis is not based on Y's intervened inadequacy (depending on his own features, or on the new conflicting context, or on the presence of other more efficient and valued competitors). But the problem is that changing the priorities among X*'s* goals, also changes the task X has to delegate and maybe Y has not got the right features for this new task (at least in X's beliefs).

This last example shows how our trust model is able to reconcile the two main cognitive ingredients: *beliefs* and *goals*. On this basis it can produce relevant forecasts: trust can change or collapse on both the brows (and they are very different phenomena). Current models (expecially in social, political and economical fields) neglect the relevant role of the *goals*, superficially disregarding the implications of a deep analysis between beliefs and goals differences.

Resuming and concluding this paragraph on the trust crisis we would like underline this difference:

a) On the one hand, there could be *beliefs crisis* (*revision*): change of opinion, reception of new information and evidences about Y's features, abilities, virtues, willingness, persistence, honesty, loyalty, and so on. As a consequence X's evaluation and trust can collapse.
b) Very different, on the other hand, is the *crisis of goals*: if X no longer has that goal, she does not want achieve it, the crisis of trust between X and Y is very different. It has in fact concluded the presupposition, the assumption and the willingness of the cooperation, of the delegation.

X does not think about being dependent on Y: she is not more interested about what Y does or is able to do: it is irrelevant for her. The detachment is more basic and radical: referring to

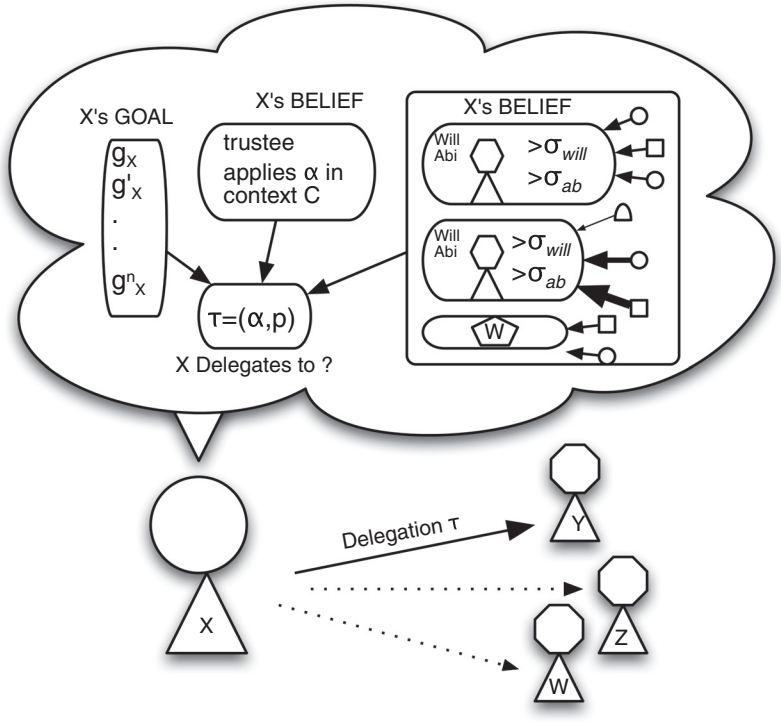

Figure 6.20 *X*'s Mental Ingredients for Delegating a task (considering other potential trustees, and different goals)

our model is the main network (the *dependence network* in our terms), that is falling, not only the (also very but less important) trust network (see also section 6.7.1).

6.8 Concluding Remarks

Strategies and devices for trust building should take into account the fact that social trust is a very dynamic phenomenon both in the mind of the agents and in society; not only because it 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. We have in fact explained how *trust is influenced by trust* in several rather complex ways. In particular we have discussed three crucial aspects of such a dynamics and we have characterized some mechanism responsible for it and some preliminary formalization of it. We have modelled:

a) *How A's trusting B and relying on it in situation* Ω *can actually (objectively) influence B's trustworthiness in* Ω. 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. Both *B*'s ability (for example through *B*'s increased self-confidence) and *B*'s willingness and disposition can be affected by *A*'s trust or distrust and delegation. *B* can for example

accept *A*'s tacit exploitation and adopt *A*'s goal, or negatively react to that. *A* can be aware of and take into account the effect of its own decision in the very moment of that decision. This make the decision of social trusting more complicated and 'circular' than a trivial decision of relying or not on a chair.

b) *How trust creates a reciprocal trust, and distrust elicits distrust*; for example because B knows that A's is now dependent on it and it can sanction A in case of violation; or because B believes that bad agents are suspicious and diffident (attributing to the other similar bad intentions) and it interprets A's trust as a sign of lack of malevolence. We also argued that the opposite is true: *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*.

c) *How diffuse trust diffuses trust* (*trust atmosphere*), that is how *A*'s trusting *B* can influence *C* trusting *B* or *D*, and so on. Usually, this is a macro-level phenomenon and the individual agent does not calculate it. We focused on *pseudo-transitivity* arguing how indirected or mediated trust always depends on trust in the mediating agent: my accepting X's evaluation about Z or X's reporting of Z's information depends on my evaluation of Z's reliability as evaluator and reporter. In other words this is not a logical or automatic process, but it is cognitively mediated.

We also discussed a more basic form of trust contagion simply due to diffusion of behaviours and to imitation because of a feeling of safety. Usually, these are macro-level phenomena and the individual agents do not calculate it.

d) *How trust can be transferred among agents on the basis of generalization of both tasks and agent's features*, that is how it is possible to predict how/when an agent who trusts something/someone will therefore trust something/someone else, before and without a direct experience. It should be clear that any theory of trust just based on or reduced to a probability index or a simple measure of experience and frequency (of success and failure) cannot account for this crucial phenomenon of a principled trust transfer from one agent to another or from one task to another. Only an explicit attributional model of the 'qualities' of the trustee that make her 'able' and 'willing' to (in the trustor's opinion), and of the 'requirements' of τ as related to the trustee's qualities, can provide such a theory in a principled way.

Our cognitive model of trust, with its analytical power, seems able to account for the inferential generalization of trustworthiness from task to task and from agent to agent not just based on specific experience and/or learning. It should also be clear how important and how productive this way of generating and propagating trust beyond direct experience and reputation should be.

References

Baier, A. (1994) Trust and its vulnerabilities in *Moral Prejudices*, Cambridge, MA: Harvard University Press, pp. 130–151.

Barber, S. and Kim, J. (2000) Belief Revision Process based on trust: agents evaluating reputation of information sources, *Autonomous Agents 2000 Workshop on 'Deception, Fraud and Trust in Agent Societies'*, Barcelona, Spain, June 4, pp. 15–26.

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

Bratman, M. E. (1987) *Intentions, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Castelfranchi, C. (1991) Social Power: a missed point in DAI, MA and HCI. In *Decentralized AI*. Y. Demazeau & J.P. Mueller (eds.) (Elsevier, Amsterdam) pp. 49–62.

Castelfranchi, C. Reasons: belief support and goal dynamics. *Mathware & Soft Computing*, 3: 233–247, 1996.

Castelfranchi, C. and Falcone, R. (1998) Towards a theory of delegation for agent-based systems, *Robotics and Autonomous Systems*, SI on Multi-Agent Rationality, Elsevier Editor, 24(3-4): 141–157.

Castelfranchi, C. and Falcone, R. (1998) Principles of trust for MAS: cognitive anatomy, social importance, and quantification. Proceedings of the International Conference on Multi-Agent Systems (ICMAS'98).

Cohen, P. and Levesque, H. (1987) Rational interaction as the basis for communication. Technical Report, N°89, CSLI, Stanford.

Conte, R. and Paolucci, M. (2002) *Reputation in Artificial Societies. Social Beliefs for Social Order*. Boston: Kluwer Academic Publishers.

Falcone, R. and 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.

Festinger, L. (1957) *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Gambetta D. (1998) Can we trust trust? In *Trust: Making and Breaking Cooperative Relations*, edited by D. Gambetta. Oxford: Blackwell.

Gambetta, D. (ed.) (1990) *Trust*. Basil Blackwell, Oxford.

Jennings. N. R. Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review*, 3: 223–250, 1993.

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

Jurca, R and Faltings, B. (2003) Towards incentive-compatible reputation management, In R. Falcone, S. Barber, L. Korba and M. Singh (eds.) *Trust, Reputation, and Security: Theories and Practice*, LNAI 2631 Springer. pp. 138–147.

Sabater, J. and Sierra, C. (2001) Regret: a reputation model for gregarious societies. In Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems, pp. 475–482. ACM Press, NewYork.

Sichman, J. R. Conte, C. Castelfranchi, Y., Demazeau (1994) A social reasoning mechanism based on dependence networks. In *Proceedings of the 11th ECAI*.

Weiner, B. (1992) *Human Motivation: Metaphors, Theories and Research*, Newbury Park, CA: Sage Publications.

M. Witkowski, A. Artikis, and J. Pitt (2001) Experiments in building experiental trust in a society of objective-trust based agents. In R. Falcone, M. Singh, and Y. Tan (eds.) *Trust in Cyber-societies: Integrating the Human and Artificial Perspectives*, LNAI 2246 Springer. pp. 111–132.