Tuning the Agent Autonomy: the Relationships between Trust and Control[1]

Rino Falcone and Cristiano Castelfranchi
CNR - Institute of Cognitive Sciences and Technologies
Rome, Italy
{castel,falcone}@ip.rm.cnr.it

Abstract

The relationship between trust and control is quite relevant both for the very notion of trust and for modelling and implementing trust-control relations with autonomous systems. We claim that control is antagonistic of the strict form of trust: "trust in y"; but also that it completes and complements it for arriving to a global trust. In other words, putting control and guaranties is trust-building; it produces a sufficient trust, when trust in y's autonomous willingness and competence would not be enough. We also argue that control requires new forms of trust: trust in the control itself or in the controller, trust in y as for being monitored and controlled; trust in possible authorities; etc. Finally, we show that paradoxically control could not be antagonistic of strict trust in y, but it can even create, increase it by making y more willing or more effective. In conclusion, depending on the circumstances, control makes y more reliable or less reliable; control can either decrease or increase trust. A good theory of trust cannot be complete without a theory of control.

1 Introduction: to trust or to control? Two opposite notions and parties

The relation between trust and control is very important and perhaps even definitory; however it is everything but obvious and linear.

On the one side, some definitions delimite trust precisely thanks to control as its opposite. But it is also true that control and guaranties make me more confident when I do not have enough trust in my partner: and what is confidence if not a broader form of trust?

On the other side, it appears that the "alternative" between control and trust is one of the main *tradeoff* in several domains of IT and computer science, from HCI to MAS, EC, virtual organisations, and so on, precisely like in human social interaction.

Consider for example the problem to mediate between two diverging concepts as control and autonomy (and the trust on which the autonomy is based) in the design of human-computer interfaces (Hendler, 1999): "One of the more contentious issues in the design of human-computer interfaces arises from the contrast between 'direct manipulation' interfaces and autonomous agent-based systems. The proponents of direct manipulation argue that a human should always be in control - steering an agent should be like steering a car - you're there and you're active the whole time. However, if the software simply provides the interface to, for example, an airlines booking facility, the user must keep all needs, constraints and preferences in his or her own head. (...) A truly effective internet agent needs to be able to work for the user when the user isn't directly in control."

Consider also the naive approach to security and reliability in computer mediated interaction, just based on strict rules, authorisation, cryptography, inspection, control, etc. (Castelfranchi, 2000) which can be in fact self-defeating for improving EC, virtual organisation, cyber-communities (Nissenbaum, 1999).

The problem is that the trust-control relationship is both conceptually and practically quite complex and dialectic. We will try to explain it both at the conceptual and modelling level, and in terms of their reciprocal dynamics.

2 What trust is: a cognitive approach

Let us recapitulate our cognitive approach and definition of trust (Castelfranchi and Falcone, 1998; Castelfranchi and Falcone, 2000).

The word "trust" is ambiguous: it denotes both the simple trustor's evaluation of trustee before relying on it (we will call this "core trust"), the same plus the decision of

[1] This paper has been partially developed within the European Project ALFEBIITE: IST-1999-10298: and in part by the TICCA Project: joint research venture between the Italian National Research Council -CNR- and Provincia Autonoma di Trento.

relying on trustee (we will call this part of the complex mental state of trust "reliance"), and the *action* of trusting, depending upon trustee (this meaning really overlaps with "delegation" (Castelfranchi and Falcone, 1998-b) and we will not use the term "trust" for this).
In Fig.1 we show how these three steps of the trust concept are causally related. In fact, there may be several evaluations of other agents (y or z) about a given task (τ); each of these evaluations is based on various parametres/components (see below); the match among these evaluations permits to decide if and which agent rely on. We should consider also external constraints that could influence our preferences/conveniences and then this decision (for example an obligation to take a decision even if nobody has had a good evaluation). Then, the decision permits to make (or not) a delegation action.

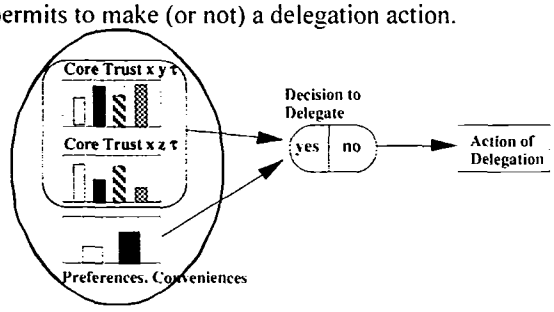

Fig.1: The decision to trust: comparing 2 trustees

Trust is, first of all, *a mental state*, an *attitude* towards an other agent (usually a social attitude). We will use the three argument predicate -**Trust(x y τ)** (where x and y are the trustor and the trustee respectively and τ=(α,g) is the task. the pair action-goal) - to denote a specific *mental state* compound of other more elementary mental attitudes (beliefs. goals. etc.). While we use a predicate **Delegate(x y τ)** to denote the *action* and the resulting relation between x and y .
Delegation necessarily is an *action.* a result of a decision. and it also creates and is a *(social) relation* among x. y. and τ. The external, observable action/behaviour of delegating either consists of the action of provoking the desired behaviour, of convincing and negotiating, of charging and empowering, or just consists of the action of doing nothing (omission) waiting for and exploiting the behaviour of the other. Indeed, will we use trust and reliance only to denote the mental state preparing and underlying delegation (*trust* will be both: the small nucleus and the whole).
Trust is normally *necessary* for delegation, but it is not *sufficient*: delegation requires a richer decision that contemplates also conveniences and preferences. As a state. trust is the most important part of the mental counter-part of delegation. i.e. that it is a structured set of mental attitudes characterising the mind of a delegating agent/trustor.
The decision to delegate has no degrees: either x delegates or x does not delegate. Indeed trust has degrees: x trusts y more or less relatively to τ. And there is a threshold under which trust is not enough for delegating.

2.1 Beliefs in Trust

We start identifying the basic ingredients of the mental state of trust.
To have trust it is necessary that the trustor has got a goal. In fact, x has a **goal** g that x tries to achieve by using y: This is what x would like to "delegate to" y, its task.
In addition, x has some specific basic beliefs:
1. "Competence" Belief: a *positive evaluation* of y is necessary, x should believe that y is useful for this goal of its, that y can produce/provide the expected result, that y can play such a role in x's plan/action, that y has some function.
2. "Disposition" Belief: Moreover, x should think that y not only is able and can do that action/task, but y actually will do what x needs. With cognitive agents this will be a belief relative to their *willingness*: this make them predictable.
These are the two prototypical components of trust as an attitude towards y. They will be enriched and supported by other beliefs depending on different kind of delegation and different kind of agents; however they are the real cognitive kernel of trust.
The kernel ingredients we have just identified are not enough for arriving to a delegation or reliance disposition. At least a third belief is necessary for this:
3. Dependence Belief: x believes -to trust y and delegate to it- that either x needs it, x depends on it (*strong dependence* (Sichman et al., 1994), or at least that it is better to x to rely rather than do not rely on it (*weak dependence* (Jennings, 1993)).
In other terms, when x trusts on someone, x is in *a strategic situation* (Deutsch, 1973): x believes that there is interference (Castelfranchi, 1998) and that its rewards, the results of its projects, depend on the actions of another agent y.
These beliefs (plus the goal g) define its "trusting y" or its **"trust in y"** in delegation. However, another crucial belief arises in x's mental state -supported and implied by the previous ones:
4. Fulfilment Belief: x believes that g will be achieved (thanks to y in this case). This is the **"trust that" g.**
Thus, *when x trusts y for g, it has also some trust that g.*
When x decides to trust, x has also the new goal that y performs α, and x rely on y's α in its plan (delegation). In other words, on the basis of those beliefs about y, x "leans against", "count on", "depends upon", "relies on", in other words x practically "trusts" y. Where -notice- "to trust" does not only means those basic beliefs (the core) but also the

decision (the broad mental state) and the act of delegating (see Fig.1).

To be more explicit: *on the basis of those beliefs about y, x decides of not renouncing to g, not personally bringing it about, not searching for alternatives to y, and to pursue g through y.*

When applied to a cognitive, intentional agent, the "Disposition Belief" must be articulated in and supported by a couple of other beliefs:

2a. Willingness Belief: *x* believes that *y* has decided and intends to do α. In fact for this kind of agent to do something, it must intend to do it. So trust requires modelling the mind of the other.

2b. Persistence Belief: *x* should also believe that *y* is stable enough in its intentions, that has no serious conflicts about α (otherwise it might change its mind), or that *y* is not unpredictable by character, etc.

2.5 Internal (trustworthiness) versus external attribution of trust

We should also distinguish between trust 'in' someone or something that has to act and produce a given performance thanks to its *internal* characteristics, and the global trust in the global event or process and its result which is also affected by external factors like opportunities and interferences (see Fig.2).

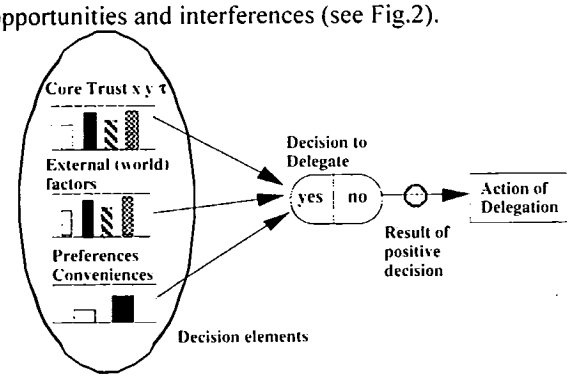

Fig.2: The decision to trust: internal and external factors

Trust *in y* (for example, 'social trust' in strict sense) seems to consists in the two first prototypical beliefs/evaluations we identified as the basis for reliance: *ability/competence* (that with cognitive agents includes self-confidence), and *disposition* (that with cognitive agents is based on willingness, persistence, engagement, etc.). Evaluation about *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.

We will call this part of the global trust (the trust 'in' *y* relative to its internal powers - both motivational powers and competential powers) *internal trust* or subjective *trustworthiness*. In fact this trust is based on an 'internal causal attribution' (to *y*) on the causal factors/probabilities of the successful or unsuccessful event.

Trust can be said to consist of or better to (either implicitly or explicitly) imply the *subjective probability* of the successful performance of a give behaviour α, and it is on the basis of this subjective perception/evaluation of risk and opportunity that the agent decides to rely or not, to bet or not on *y*. However, the probability index is based on, derives from those beliefs and evaluations. In other terms the global, final probability of the realisation of the goal *g*, i.e. of the successful performance of α, should be decomposed into the probability of *y* performing the action well (that derives from the probability of willingness, persistence, engagement, competence: *internal attribution*) and the probability of having the appropriate conditions (opportunities and resources *external attribution*) for the performance and for its success, and of not having interferences and adversities (*external attribution*).

Strategies to establish or incrementing trust are very different depending on the external or internal attribution of your diagnosis of lack of trust. If there are adverse environmental or situational conditions your intervention will be in establishing protection conditions and guarantees, in preventing interferences and obstacles, in establishing rules and infrastructures; while if you want to increase your *trust in* your contractor you should work on its motivation, beliefs and disposition towards you, or on its competence, self-confidence, etc..

Environmental and situational trust (which are claimed to be so crucial in electronic commerce and computer mediated interaction; see for ex. (Rea, 2000) are aspects of the external trust. It is important to stress that: *when the environment and the specific circumstances are safe and reliable, less trust in y is necessary for delegation.*

Vice versa, when *x* strongly trusts *y*, his capacities, willingness and faithfulness, *x* can accept a less safe and reliable environment. We account for this 'complementariety' between the internal and the external components of trust in *y* for *g* in given circumstances and a given environment.

However, as we will see later we shouldn't identify 'trust' with 'internal or interpersonal or social trust' and claim that when trust is not there, there is something that can replace it (ex. surveillance, contracts, etc.). It is just matter of different kinds or better *facets of trust.*

3 What control is

The control is a (meta) action:

a) aimed at ascertaining whether another action has been successfully executed or if a given state of the world has been realized or maintained (*feedback, checking*);

b) aimed at dealing with the possible deviations and unforeseen events in order to positively cope with them (*intervention*).

When the client is delegating a given object-action, what about its control actions? Considering, for the sake of simplicity, that the control action is executed by a single agent, when Delegates(Ag_1 Ag_2 τ) there are at least four possibilities:

i) Ag_1 delegates the control to Ag_2: the client does not (directly) verify the success of the delegated action to the contractor;

ii) Ag_1 delegates the control to a third agent;

iii) Ag_1 gives up the control: nobody is delegated to control the success of α;

iv) Ag_1 maintains the control for itself.

Each of these possibilities could be explicit or implicit in the delegation of the action, in the roles of the agents (if they are part of a social structure), in the preceding interactions between the client and contractor, etc.

To understand the origin and the functionality of control it is necessary to consider that Ag_1 can adjust run-time its delegation to Ag_2 if it is in condition of:

a) receiving in time the necessary information about Ag_2's performance (*feedback*);

b) intervening on Ag_2's performance to change it before its completion (*intervention*).

In other words, Ag_1 must have some form of "control" on and during Ag_2's task realisation.

Control requires feedback plus intervention (Fig.3).

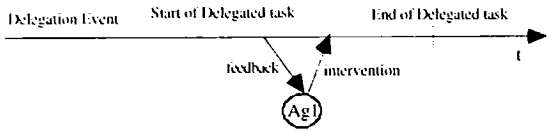

Fig.3: Control channels for the client's adjustment

Otherwise no adjustment is possible. Obviously, the feedback useful for a run-time adjustment must be provided timely for the intervention. In general, the feedback activity is the precondition for an intervention; however it is also possible that either only the feedback or only the intervention hold.

Feedback can be provided by observation of Ag_2's activity (inspection, surveilance, monitoring), or by regularly sent messages by Ag_2 to Ag_1, or by the fact that Ag_1 receives or observes the results/products of Ag_2's activity or their consequences.

As for *Intervention* we consider five kinds of intervention:

- *stopping the task* (the delegation or the adoption process is suddenly interrupted);
- *substitution* (an intervention allocates part of the (or the whole) task to the intervening agent);
- *correction of delegation* (after the intervention, the task is partially or totally changed);
- *specification or abstraction of delegation* (after the intervention, the task is more or less constrained);
- *repairing of delegation* (the intervention leaves the task activity unchanged but it introduces new actions necessary to achieve the goal(s) of the task itself).

Each of these interventions could be realized through either a *communication act* or a *direct action* on the task by the intervening agent.

The *frequency of the feedback on the task* could be:

- *purely temporal* (when the monitoring or the reporting is independent of the structure of the activities in the task, they only depend on a temporal choice);
- *linked with the working phases* (when the activities of the task are divided in phases and the monitoring or the reporting is connected with them).

Client and contractor could adjust the frequency of their feedback activity in three main ways:

- by *changing the temporal intervals* fixed at the start of the task delegation (in the case in which the monitoring/reporting was purely temporal);
- by *changing the task phases* in which the monitoring/reporting is realized with respect to those fixed at the start of the task delegation (in the case in which monitoring/reporting was linked with the working phases);

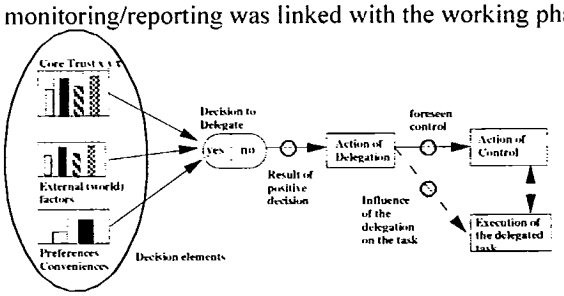

Fig.4: Decision, delegation and control

- by *moving from* the purely temporal monitoring/reporting to the working phases monitoring/reporting (or vice versa).

Also the *frequency of intervention* is relevant. As explained above, the intervention is strictly connected with the presence of the monitoring/reporting on the task, even if, in principle, both the intervention and the monitoring/reporting could be independently realized. In addition, also the frequencies of intervention and of monitoring/reporting are correlated. More precisely, the frequency of intervention could be:

1) *never*; 2) *just sometimes* (phase or time, a special case of this is at the end of the task); 3) *at any phase or at any time.*

Fig.4 integrates the schema of Fig.2 with the two actions: control and execution of the task.

Plans typically contain control actions of some of their actions (Castelfranchi and Falcone, 1994).

3 Control replaces trust and trust makes control superflous?

As we said before, a perspective of duality between trust and control is very frequent and at least partially valid. Consider for example this definition of trust:

"The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other party will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al., 1995)

This capture a very intuitive and common sense use of the term trust (in social interaction). In fact, it is true -in this restrict sense- that if you control me "you don't trust me!"; and it is true that if you do not trust me enough (for counting on me) you would like to monitor, control and enforce me in some way.

In this view, control and normative "remedies" "have been described as weak, impersonal substitutes for trust" (Sitkin and Roth, 1993), or as "functional equivalent... mechanisms" : "to reach a minimum level of confidence in cooperation, partners can use trust and control to complement each other" (Beamish, 1988).

With respect to this view, we have some problems:

- on the one side, it is correct, it captures something important. However, in such a complementariety, how the control precisely succeeds in augmenting confidence, is not really modelled and explained.
- on the other side, there is something reductive and misleading in such a position:

- it reduces trust to a strict notion and looses some important uses and relations:
- it ignores different and additional aspects of trust also in the trustee;
- it misses the point of considering control as a way of increasing the strict trust in the trustee.

We will argue that: firstly, control is antagonistic to strict trust; secondly, it requires new forms of trust and build the broad trust; thirdly, it completes and complements it; finally, it can even create, increase the strict trust. As you can see a quite complex relationship.

4.1 A strict trust notion (antagonist of control) and a broad notion (including control)

As we said we agree on the idea that (at some level) trust and control are antagonistic (one eliminates the other) but complementary. We just consider this notion of trust -as defined by Mayer- too restricted.

It represents for us the notion of trust in strict sense, i.e. applied to the agent (and in particular to a social agent and to a process or action), and strictly relative to the "internal attribution", to the internal factor. In other words, this represent the "trust in y" (as for action α and goal g). But this trust -when is enough for delegation- implies the "trust that" (g will be achieved or maintained); and anyway it is part of a broader trust (or non-trust) that g.

We consider both forms of trust. Also the trust (or confidence) in y, is, in fact, just the trust (expectation) that y is able and will appropriately do the action α (that I expect for its result g). But the problem is: are such an ability and willingness (the "internal" factors) enough for realizing g? What about conditions for successfully executing α (i.e. the opportunities)? What about other concurrent causes (forces, actions, causal process consequent to y's action)? If my trust is enough for delegating to y, this means that I expect, trust that g will probably be realized.

We propose a broader notion of trust including all my expectations (about y and the world) such that g will be eventually true thanks (also) to y's action; and a strict notion of trust as "trust in" y, relative only to the internal factors (see Fig.5).

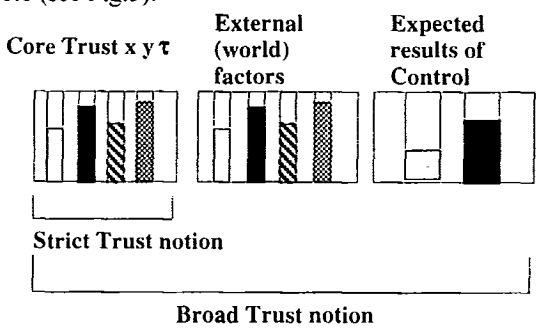

Fig.5: Control complements strict trust

This strict notion is similar to which defined by Mayer (apart from the lack of the competence ingredient), and it is in contrast, in conflict with the notion of control. If there is control then there is no trust. But on the other side they are also two complementary parts, as for the broad/global trust: control supplements trust.

In this model, trust in y and control of y are *antagonistic*: where there is trust there is no control, and viceversa; the larger the trust the less room for control, and viceversa; but they are also *supplementary*: one remedies to the lack of the other; they are parts of one and the same entity. In this perspective notice that control is both antagonist to (one form of trust) and consituent of (another form of) trust.

Obviously, this schema is very simplistic and just intuitive. We will make this idea more precise. However let us remarke immediately that this is not the only relation between strict-trust and control. Control is not only aimed at supplementing and "completing" trust (when trust in y would not be enough); it can be also aimed precisely at augmenting the internal trust in y, y's trustwortiness.

4.2 Relying on control and bonds requires additional trust

To our account of trust one might object that we overstate the importance of trust in social actions such as contracting, and organisations; since everything is based on delegation and delegation presupposes enough trust. In fact, it might be argued -within the duality framework- that people put contracts in place precisely because they do *not* trust the agents they delegate tasks to. Since there is no trust people want to be protected by the contract. The key in these cases would not be trust but the ability of some authority to assess contract violations and to punish the violators. Analogously, in organisations people would not rely on trust but on authorisation, permission, obligations and so forth.

In our view this opposition is fallacious: it seems that trust is only relative to the character or friendliness, etc. of the trustee. In fact in these cases (control, contracts, organisations) we just deal with *a more complex and specific kind of trust.* But trust is always crucial.

We put control in place only because we believe that the trustee will not avoid or trick monitoring, will accept possible interventions, will be positively influenced by control. We put a contract in place only because we believe that the trustee will not violate the contract, etc.. These beliefs are nothing but "trust".

Moreover, when true contracts and norms are there, this control-based confidence require also that x *trusts* some authority or its own ability to monitor and to sanction y (see (Castelfranchi and Falcone, 1998-b)) on *three party trust*). x must also trust procedures and means for control (or the agent delegated to this task).

4.3 How control increases and complements trust

As we saw, control in a sense complements and surrogates trust and makes broad trust notion (see Fig.5) sufficient for delegation and betting. How does this work? How does control precisely succeed in augmenting confidence?

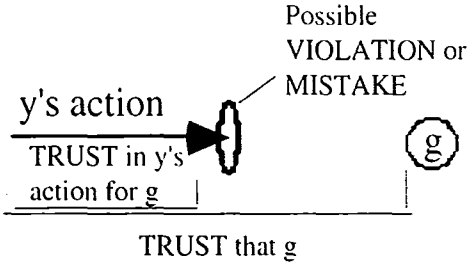

Fig.6: Trust in the action Vs trust in the result

Our basic idea, is that strict-trust (trust in y) is not the complete scenario: to arrive from the belief that "Brings y about that action α" (it is able and willing, etc.) to the belief that "eventually g", something is lacking: the other component of the global trust: more precisely, the trust in the "environment" (external conditions), including the intervention of the trustor or of somebody else. Control can be aimed at filling this gap between y's intention and action and the desired result "that g" (Fig.6).However, does control augment only the broad trust? Not true: the relationship is more dialectic. It depends on the kind and aim of control. In fact, it is important to understand that trust (also trust in y) is not a ante-hoc and static datum (either sufficient or insufficient for delegation before the decision to delegate). It is a dynamic entity; for example there are effects, feedback of the decision to delegate on its own pre-condition of trusting y. Analogously the decision to put control can affect the strict-trust whose level make control necessary!

Thus the schema: trust+control, is rather simplistic, static, a-dialectic; since the presence of control can modify and affect the other parameters. There are indeed two kinds and functions of control.

4.3.1 Two kinds of Control

A) *Pushing or influencing control*: preventing violations or mistakes

The first kind or function of control is aimed at operating on the "trust in y" and more precisely at increasing it. It is aimed in fact at reducing the probability of y's defaillance, slips, mistakes, deviations or violation; i.e., at preventing and avoiding them. The theory behind this kind of surveillance at least one on the following beliefs:

i) if y is (knows to be) surveilled its performance will be better because either it will put more attention, or more effort, or more care, etc. in the execution of the delegated task; in other words, *it will do the task better;* or

ii) if y is (knows to be) surveilled it will be more reliable, more faithful to its commitment, less prone to violation; in other words, *it more probably will do the task* .

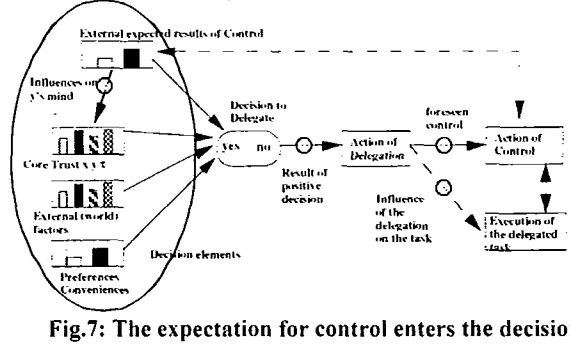

Fig.7: The expectation for control enters the decision of trusting

Since x believes this, by deciding of controlling y (and letting y knows about this) x increases its own evaluation/expectation (i.e., its trust) about y's willingness, persistence, and quality of work. As we can see in Fig.7, one of the control results is just to change the core trust of x on y about τ. This form of control is essentially *monitoring*

(inspection, surveillance, reporting, etc.), and can work also without any possibility of *intervention*. Indeed, *it necessarily requires that y knows about being surveilled.* This can be just a form of 'implicit communication' (to let the other see/believe that we can see him, and that we know that he knows, etc.), but frequently the possibility of some explicit communication on this is useful ("don't forget that I see you!"). Thus, also some form of *intervention* can be necessary: a communication channel.

B) *Safety, correction or adjustment control*: preventing failure or damages

This control is aimed at preventing dangers due to y's violations or mistakes, and more in general is aimed at having the possibility of adjustment of delegation and autonomy of any type (Falcone and Castelfranchi, 2000-b). In other words, it is not only for repairing but for correction, through advises, new instructions and specifications, changing or revoking task, direct reparation, recover, or help, etc.

For this reason this kind of control is possible only if some intervention is allowed, and requires monitoring (feedback) run-time.

This distinction is close to the distinction between "control for prevention" and "control for detection" used by (Bons et al., 1998). However, they mainly refer to legal aspects of contracts, and in general to violations. Our distinction is related to the general theory of action (the function of control actions) and delegation, and is more general. The first form/finality of control is preventive not only of violations (in case of norms, commitments, or contracts) but also of missed execution or mistakes (also in weak delegation where there are no obligations at all). The second form/finality is not only for sanctions or claims, but for timely intervening and preventing additional damages, or remedying and correcting (thus also the second can be for prevention, but of the consequences of violation). "Detection" is just a means; the real aim is intervention for safety, enforcement or compensation.

Moreover, we argue that an effect (and a function/aim) of the second form of control can be also to prevent violation: this happens when the controlled agent knows or believes - before or during his performance - that there will be "control for detection" and worries about this (sanctions, reputation, lack of autonomy, etc.).

4.4 Filling the gap between doing/action and achieving/results

Let's put the problem in another perspective. As we said, trust is the background for delegation and reliance i.e., to "trust" as a decision and an action; and it is instrumental to the satisfaction of some goal. Thus the trust in y (sufficient for delegation) implies the trust that g (the goal for which x counts on y) will be achieved.

Given this two components or two logical step scenario, we can say that the first kind of control is pointing to, is impinging on the first step and is aimed at increase it; while the second kind of control is pointing to the second step and is aimed at increasing it, by making more sure the achievement of g also in case of defaillance of y.

In this way the control (monitoring plus intervention) complement the trust in y which would be insufficient for achieving g, and for delegating; this additional assurance (the possibility to correct work in progress y's activity) makes x possible to delegate to y g. In fact in this case x is not only count on y, but x counts on a multi-agent possible plan that include possible actions of its.

As we can see from the formula (a) the important thingh is that y believes that the control holds, and not if it is really holds. For example, x could not trust enough y and communicate to it the control: this event modifies the y's mind and the x's judge about trusting y.

Thus, in trust-reliance without the possibility of intervention for correction and adjustment, there is only one possibility for achieving g, and one activity (y's activity) x bets on (Fig.8).

Fig.8: The gap betwwen action and expected results

While, if there is control for adjustment, the achievement of g is committed to y's action plus x's possible action (intervention), x bets on this combination (Fig.9).

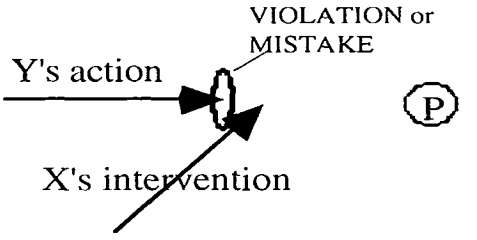

Fig.9: Intervention in the gap

A very similar complementing or remedying role are guaranties, protections, and assurance. I do not trust enough the action, and I put protections in place to be sure about the desired results. For example, I do not trust to drive a motorcicle without a crash-helmet, but I trust to do so with it.

4.5 The dynamics

It is important underlaining that the first form/aim of control is oriented at increasing the *reliability* of y (in terms of fidelity, willingness, keeping promises, or in terms of

carfulnees, concentration and attention) and then it is a way of increasing x's trust in y which should be a presupposition not an effect of my decision: x believes that (if x surveilles y) y will be more committed, willing and reliable; i.e. the strenght of x's trust-beliefs in y and thus x's degree of trust in y are improved.

This is a very interesting social (moral and pedagogigal) strategy. In fact it is in opposition with another well know strategy aimed at increasing y's trustworthiness; i.e., "trust creates trust"!

In fact, precisely the reduction/renounce to control is a strategy of "responsabilisation" of y, aimed at making it more reliable, more committed.

Those stategies are in conflict with each other. When and why do we chose to make y more reliable and trustworty through responsabilization (renounce to surveillance), and when through surveillance? A detailed model of how and why trust creates/increases trust is necessary to answer this question.

Should we make our autonomous agents (or our cyber-partners) more reliable and trustworty through responsabilization or through surveillance?

We will not have this doubt with artificial agents, since their "psychology" will be very simple and their effects will not be very dynamic. At least for the moment with artificial agents control will complement insufficient trust and perhaps (known control) will increase commitment. However, for sure those subtle intertaction problems will be relevant for computer mediated human interaction and collaboration.

5 Conclusions

As we saw relationships between trust and control are rather complicated. On the one side, it is true that where/when there is trust there is no control, and vice versa. But this is a restricted notion of trust: it is "trust in y", which is just a part, a component of the whole trust needed for relying on the action of another agent. Thus we claimed that control is antagonistic of this strict form of trust: but also that it completes and complements it for arriving to a global trust. We have also argued that control requires new forms of trust: trust in the control itself or in the controller, trust in y as for being monitored and controlled; trust in possible authorities; etc.

Finally, we have shown that paradoxically control could not be antagonistic of strict trust in y, but it could even create, increase the trust in y, making y more willing or more effective. In conclusion, depending on the circumstances, control makes y more reliable or less reliable.

6 References

Beamish P., (1988). Multinational joint ventures in developing countries. London: Routledge.

Bons R., Dignum F., Lee R., Tan Y.H., (1998). A formal specification of automated auditing of trustworthy trade procedures for open electronic commerce. *Autonomous Agents '99 Workshop on "Deception, Fraud and Trust in Agent Societes"*, Minneapolis, USA, May 9, pp.21-34.

Castelfranchi, C., (2000). Formalizing the informal? Invited talk DEON2000 Toulose.

Castelfranchi C., 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)*, Paris, July, pp.72-79.

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

Castelfranchi, C., Falcone, R. (1994). Towards a theory of single-agent into multi-agent plan transformation. The *3rd Pacific Rim International Conference on Artificial Intelligence (PRICAI94)*, Beijing, China, 16-18 agosto, pp.31-37.

Castelfranchi C., (1998). Modelling Social Action for AI Agents. *Artificial Intelligence*, 103, pp. 157-182.

Deutsch M., The Resolution of Conflict. Yale University Press, New Haven and London, 1973.

Falcone R., Castelfranchi C., (2001). Social Trust: A Cognitive Approach, in *Trust and Deception in Virtual Societies* by Castelfranchi C. and Yao-Hua Tan (eds), Kluwer Academic Publishers, pp. 55-90.

Falcone, R., Castelfranchi, C., (2000-b). Levels of Delegation and Levels of Adoption as the basis for Adjustable Autonomy. *Lecture Notes in Artificial Intelligence* n°1792, pp.285-296.

Hendler J., (1999). Is there an Intelligente Agent in your Future?, http://helix.nature.com/webmatters/agents/agents.html

Jennings, N.R. (1993). Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review*, 3, 223-50.

Nissenbaum H., (1999). Can trust be Secured Online? A theoretical perspective; http://www.univ.trieste.it/~diptfilo/etica_e_politica/1999_2/nissenbaum.html

Mayer R.C., Davis J.H., Schoorman F.D., (1995). An integrative model of organizational trust, Academy of Management Review, Vol.20, N°3, pp. 709-734.

Rea Tim, (2000). Engendering Trust in Electronic Environments - Roles for a Trusted Third Party: in *Deception, Fraud and Trust in Virtual Societies* by Castelfranchi C. and Yao-Hua Tan (eds), Kluwer Academic Publisher, pp.221-234.

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

Sitkin S.B., and Roth N.L. (1993). Explaining the limited effectiveness of legalistic "remedies" for trust/distrust. Organization Science, Vol.4, pp.367-392.