Applied Artificial Intelligence, 14:799–823, 2000

0883-9514/00 $12.00 + .00

TRUST AND CONTROL: A DIALECTIC LINK

CRISTIANO CASTELFRANCHI AND RINO FALCONE
National Research Council-Institute of Psychology, Unit of "AI, Cognitive Modelling, and Interaction," Roma, Italy

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, but it is not trivial at all. On the one side, it is true that where/when there is control there is no trust, and vice versa. However, this refers to a restricted notion of trust: i.e., "trust in y," which is just a part, a component of the global trust needed for relying on the action of another agent. It is 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. In other words, putting control and guarantees is trust-building; it produces a sufficient trust, when trust in y's autonomous willingness and competence would not be enough. It is 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, it is shown that, paradoxically, control could not be antagonistic of strict trust in y, but it can even create and 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. Two kinds of control are also analyzed, characterized by two different functions: "pushing or influencing control" aimed at preventing violations or mistakes, versus "safety, correction, or adjustment control" aimed at preventing failure or damages after a violation or a mistake. A good theory of trust cannot be complete without a theory of control.

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 delimit trust precisely, thanks to control as its opposite. But it is also true that control and guarantees make one more confident when one does not have enough trust in one's partner; and what is confidence if not a broader form of trust?

This research has been supported by the IST Programme ALFEBIITE (A Logical Framework for Ethical Behaviour between Infohabitants in the Information Trading Economy of the Universal Information Ecosystem) contract IST-1999-10298.

Address correspondence to Rino Falcone, IP-CNR, Viele Marx, 15-00137, Roma, Italy. E-mail: falcone@ip.rm.cnr.it

On the other side, it appears that the "alternative" between control and trust is one of the main *tradeoffs* in several domains of IT and computer science, from human computer interaction (HCI) to multiagent systems (MAS), electronic commerce (EC), virtual organizations, and so on, precisely as 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 most 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, for example, to an airline's 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."

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

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

WHAT TRUST IS: A COGNITIVE APPROACH

Let us recapitulate the cognitive approach and definition of trust (Castelfranchi & Falcone, 1998, 2000). The word "trust" is ambiguous: it denotes both the simple trustor's evaluation of trustee before relying on it (this will be called "core trust"), the same plus the decision of relying on trustee (this part of the complex mental state of trust will be called "reliance"), and the *action* of trusting, depending upon trustee (this meaning really overlaps with "delegation" (Castelfranchi & Falcone, 1998b) and the term "trust" will not be used for this).

In Figure 1, it will be shown how these three steps of the trust concept are causally related. In fact, there may be several evaluations of other agents (y or z in the figure) about a given task (τ in the figure); each of these evaluations is based on various parameters/components (see below); the match among these evaluations permits one to decide if and which agent to rely on. One should consider also external constraints that could influence

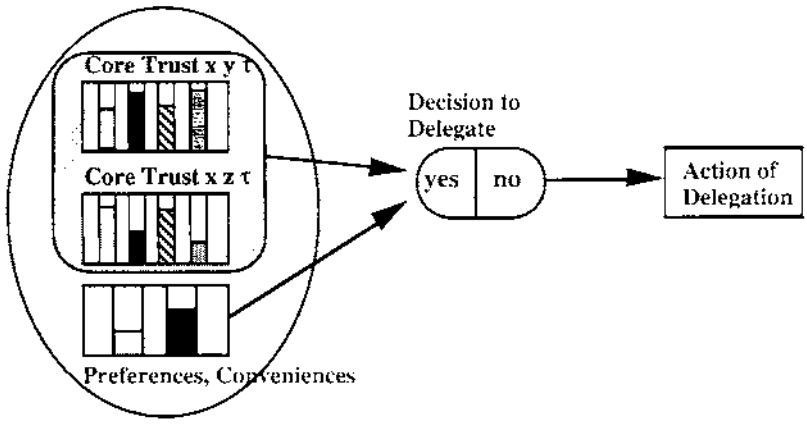

FIGURE 1. The decision to trust: Comparing two trustees.

the 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 one to make (or not) a delegation action.

Trust is, first of all, *a mental state*, an *attitude* toward another agent (usually a social attitude). The three-argument predicate -Trust(x y τ) will be used (where x and y are the trustor and the trustee, respectively, and $\tau = (\alpha,g)$ is the task, the pair action-goal) to denote a specific *mental state* compound of other more elementary mental attitudes (beliefs, goals, etc.), while one uses a predicate Delegate (x y τ) to denote the *action* and the resulting relation between x and y.

Delegation should necessarily be an *action*, the result of a decision, and it also creates and is a (*social*) *relation* among x, y, and τ. The external, observable action/behavior of delegating either consists of the action of provoking the desired behavior, convincing and negotiating, charging and empowering, or just consisting of the action of doing nothing (omission), waiting for and exploiting the behavior of the other. Indeed, trust and reliance are used only to denote the mental state preparing and underlying delegation (*trust* will be both: the small nucleus and the whole).

There may be trust without delegation: either the level of trust is not sufficient to delegate, or the level of trust will be sufficient but there are other reasons preventing delegation (for example, prohibitions, see Figure 1).

So, trust is normally *necessary* for delegation,[1] 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 counterpart of delegation, i.e., that it is a structured set of mental attitudes characterizing 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.

Basic Beliefs in Trust

The basic ingredients of the mental state of trust begin to be identified. 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 will actually do what x needs. With cognitive agents this will be a belief relative to their *willingness*: this makes them predictable.

These are the two prototypical components of trust as an attitude toward y. They will be enriched and supported by other beliefs, depending on different kind of delegation and different kinds of agents; however, they are the real cognitive kernel of trust. As will be seen later, even the goal can be varied (in negative expectation and aversive forms of "trust") but not these beliefs.

Esteem and Expectation

The nature of the above basic beliefs about y can be stressed. They are, after the decision step (see Figure 1), *evaluations* and *positive expectations*, not "neutral" beliefs. In trusting y, x believes that y has the right qualities, power, ability, competence, and disposition for g. Thus, the trust that x has in y is (clearly and importantly) part of (and is based on) its esteem, "image," and reputation (Dasgupta, 1990; Raub & Weesie, 1990).

There is also a "positive expectation" about y's power and performance. *A positive expectation is the combination of a goal and belief about the future* (*prediction*): *x believes that both g and x desires*/intends that g. In this case: x *believes* that both y can and will do; and x *desires*/wants that y can and will do.

Trust and Reliance

The kernel ingredients we have just identified are not enough for arriving at 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 to 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. Fulfillment belief: x believes that g will be achieved (thanks to y in this case).[2] 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 also has 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" not only means those basic beliefs (the core), but also the decision (the broad mental state) and the act of delegating (see Figure 1). To be more explicit: *on the basis of those beliefs about y, x decides to not renounce to g, not personally bringing it about, not searching for alternatives to y, and to pursue g through y.*

Using Meyer, van Linder, and van der Hoek's logics (Meyer & van der Hoek, 1992; van Linder, 1996), and introducing some "ad hoc" predicate (like WillDo), one can summarize and simplify the mental ingredients of trust as in Figure 2. In this figure, PE means positive expectation, B is the believe operator (the classical doxastic operator), and W is the wish operator (a normal modal operator).[3]

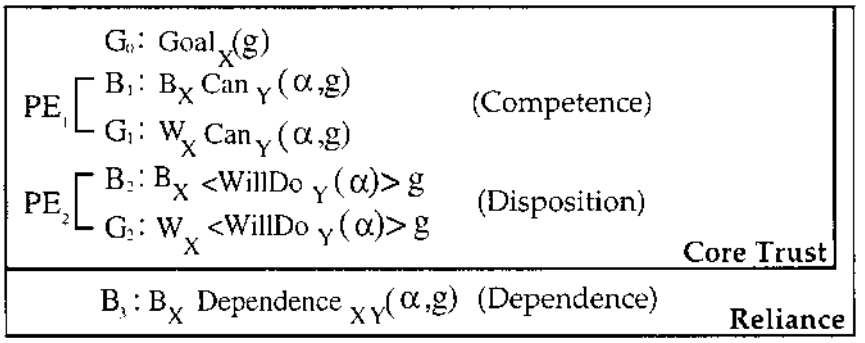

FIGURE 2. Basic mental ingredients of trust.

Wishes are the agents' desires; they model the things that the agents like to be the case; the difference between wishes and goals consists in the fact that goals are selected wishes. The fulfillment belief derives from the formulas in the above schema.

Of course, there is a coherence relation between these two aspects of trust (core and reliance): the decision of betting and wagering on y is grounded on and justified by these beliefs. More than this: the degree or strength (see later) of trust must be sufficient to decide to rely and bet on y (Marsh, 1994; Snijders & Keren, 1996). The trustful beliefs about y (core) are the presupposition of the act of trusting y.

Risk, Investment, and Bet

Any act of trusting and relying implies some bet and some risk (Luhmann, 1990). In fact, x might eventually be disappointed, deceived, and betrayed by y: x's beliefs may be wrong. At the same time x bets something on y. First, x renounced to (search for) possible alternatives (for example, other partners) and x might have lost its opportunity: thus x is risking on y the utility of its goal g (and of its whole plan). Second, x had some cost in evaluating y, in waiting for its actions, etc. and x wasted its own time and resources. Third, perhaps x had some cost to induce y to do what x wants or to have y at its disposal (for example, x paid for y or for its service); now this investment is a real bet (Deutsch, 1973) on y. Thus, to be precise we can say that: *When x trusts y there are two risks:*

a) *the risk of failure, the frustration of g* (*possibly forever, and possibly of the entire plan containing g*);[4]
b) *the risk of wasting the efforts.*

Not only x risks to miss g (*missed gains*) but x also risks to waste her investments (*loss*).

The act of trusting/reliance is a real wager, a risky activity: it logically presupposes some uncertainty, but it also requires some predictability of y, and usually some degree of trust in y. This subjective perception of risk and degree of trust can either be due to lack of knowledge, incomplete information, dynamic world, or to favorable and adverse probabilities. As we will see this makes some difference in reasons and functions for control.

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 modeling the mind of the other.

2b. Persistence belief: x should also believe that y is stable enough in its intentions, which has no serious conflicts about α (otherwise, it might change its mind), or that y is not unpredictable by character, etc.[5]

Internal (Trustworthiness) Versus External Attribution of Trust

One 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 Figure 3).

Trust *in y* (for example, "social trust" in strict sense) seems to consist in the first two prototypical beliefs/evaluations 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 the trust "in" y). An evaluation should also be added about the probability and consistence of obstacles, adversities, and interferences. One 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.

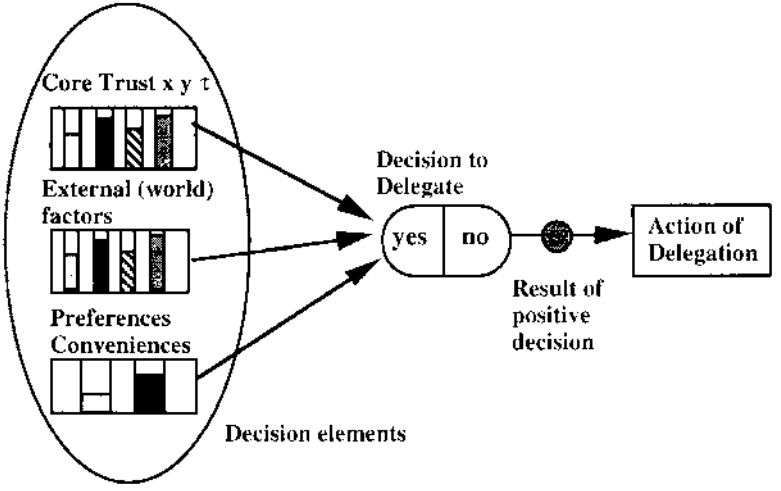

FIGURE 3. The decision to trust: Internal and external factors.

Trust can be said to consist of or better to (either implicitly or explicitly) imply the *subjective probability* of the successful performance of a given behavior 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 derivations from those beliefs and evaluations. In other terms the global, final probability of the realization 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 its success, and of not having interferences and adversities (*external attribution*).

Strategies to establish or increment 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, preventing interferences and obstacles, establishing rules and infrastructures; if you want to increase your *trust* in your contractor you should work on its motivation, beliefs, and disposition toward yourself, or on its competence, self-confidence, etc.[6]

Environmental and situational trust (which are claimed to be so crucial in electronic commerce and computer mediated interaction) (see, for example, Castelfranchi and Tan, 2000; 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 (the contractor) is necessary for delegation (for example, for transactions)*.

Vice versa, when x strongly trust y, his capacities, willingness, and faithfulness, x can accept a less safe and reliable environment (with less external monitoring and authority). This "complementarity" between the internal and the external components of trust in y for g is accounted for in given circumstances and a given environment.

However, as will be seen later, one should not identify "trust" with "internal or interpersonal or social trust" and claim that when trust is not there, there is something that can replace it (for example, surveillance, contracts, etc.). It is just a matter of different kinds or better *facets of trust*.

Formal Definition of Trust

When x relies on y for y's action, x is taking advantage of y's independent goals and intentions, predicting y's behavior on such a basis, or x is itself inducing such goals in order to exploit y's behavior. In any case, x not only believes that y is able to do and can do (opportunity), but also that y

will do because it is committed to this intention or plan (not necessarily to *x*).

Let one *simplify* and formalize this: one might characterize social trust mental state as follows:

$$\begin{aligned}\text{Trust}(X,Y,\tau) = \; & \text{Goal}_X\, g \wedge B_X \text{PracPoss}_Y(\alpha,g) \\ & \wedge\; B_X\, \text{Prefer}_X(\text{Done}_Y(\alpha,g),\, \text{Done}_X(\alpha,g)) \\ & \wedge\; (B_X(\text{Intend}_Y(\alpha,g) \wedge \text{Persist}_Y(\alpha,g)) \\ & \wedge\; (\text{Goal}_X(\text{Intend}_Y(\alpha,g) \wedge \text{Persist}_Y(\alpha,g)))\end{aligned}$$

Where: $\text{PracPoss}_Y(\alpha,g) = < \text{Do}_Y(\alpha) > g \wedge \text{Ability}_Y(\alpha)$.

In other words, trust is a set of mental attitudes characterizing the "delegating" agent's mind, which prefers another agent doing the action. *Y* is a cognitive agent, so *x* believes that *y intends to do* the action and *y will persist* in this.

Consider eventually that some kind of delegation, typical of the social sciences, (where there is agreement, promise, etc.), are, in fact, based on *y*'s awareness and implicit or explicit agreement (compliance); they presuppose *goal-adoption* by *y*. Thus, to trust *y* in this case means to trust its agreement and willingness to help/adopt (social commitment) and in its motives for doing so. This level of trust presupposes beliefs about the social mind and attitudes of the trustee.[7]

Degrees of Trust

In this model, the degree of trust of *x* in *y* is grounded in the cognitive components of *x*'s mental state of trust. More precisely, *the degree of trust* (*DoT*) *is a function of the subjective certainty of the relevant beliefs.* One uses the degree of trust to formalize a rational basis for the decision of relying and betting on *y*. Also one claims that the "quantitative" aspect of another basic ingredient is relevant: *the value or importance or utility of the goal g*. In sum, *the quantitative dimensions of trust are based on the quantitative dimensions of its cognitive constituents.*

For oneself trust is not an arbitrary index with an operational importance, without a real content, but is based on the subjective certainty of the relevant beliefs, which support each other and the decision to trust.

Trust-attitudes will be called that operator which, when applied to two agents (Ag_1 and Ag_2), a task (τ), and a temporal instant (t), returns the set of beliefs and goals (of Ag_1 on Ag_2 about τ at the time t) useful to the trust relation. One can imagine that each of the beliefs included in trust-attitudes (Ag_1 Ag_2 τ t) will have a particular weight: a degree of credibility ($\text{DoC}(B_i)$) with $0 \leq \text{DoC}(B_i) \leq 1$.

One considers here the resulting degree of trust of Ag_1 on Ag_2 about τ at time t, as the simple multiplication of all these factors (indeed one has also considered the possibility of saturation effects for each of the factors included in the multiplication and the possibility to introduce different parameters for each factor (Castelfranchi & Falcone, 2000).

If one calls *Eval-DoT* the function, which when applied to a set of mental states returns the result of the product of the weights of these mental states, one can say:

$$\text{Eval-DoT}(\text{Trust-Attitudes}(Ag_1 Ag_2\ \tau t)) = \text{DoT}_{Ag1,\,Ag2,\,\tau,\,t} \quad (0 \leq \text{DoT}_{Ag1,\,Ag2,\,\tau,\,t} \leq 1).$$

In order that Ag_1 trusts Ag_2 about τ at the time t, and then delegates that task, it is not only necessary that the $\text{DoT}_{Ag1,\,Ag2,\,\tau,\,t}$ exceeds a given (Ag_1's) threshold, but also that it constitutes the better solution (compared with the other possible solutions). So one should consider the abstract scenario[8,9] of Figure 4.

The analysis of this scenario produces one of these possible choices:

i) Ag_1 tries to achieve the goal by itself;
ii) Ag_1 delegates the achievement of that goal to another agent (Ag_2, ..., Ag_n);
iii) Ag_1 does nothing (relatively to this goal), i.e., renounces it.

It is possible to determine a trust choice starting from each combination of credibility degrees $-\{\text{DoT}_{Ag1,\,Agi,\,\tau,\,t}\}$ with $Ag_i \in \{Ag_1, \ldots, Ag_n\} -$ of the main beliefs included in trust-attitudes ($Ag_1\ Ag_i\ \tau\ t$), and from a set of Ag_1's utilities $\{U_{p+,t}, U_{p-,t}, U_{\text{di}+,t}, U_{\text{di}-,t}, U_{0,t}\} = U(Ag_1, t)$, with $i \in \{2, \ldots, n\}$.

It is possible that—once fixed the set of utilities and the kind and degree

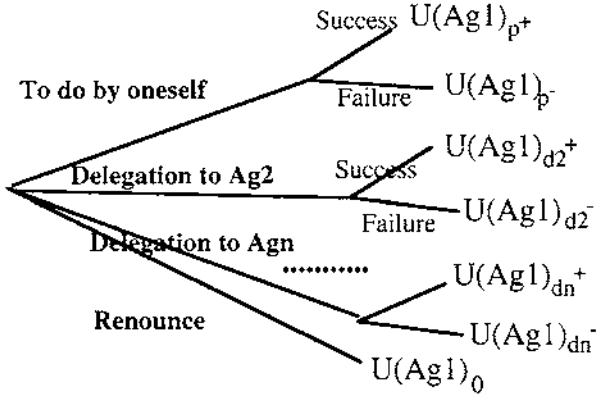

FIGURE 4. The decision scenario.

of control—different combinations of credibility degrees of the main beliefs produce the same choice. However, in general, changing the credibility degree of some beliefs more should change the final choice about the delegation (and the some holds for the utilities and the control). Obviously, at different times one could have different sets of beliefs and utilities and then a different decision about the delegation.

WHAT CONTROL IS

The control is a (meta) action aimed at: a) ascertaining whether another action has been successfully executed or if a given state of the world has been realized or maintained (feedback, checking); and b) 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 functionality of control it is necessary to consider that Ag_1 can adjust the run-time of 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 realization. *Control* requires *feedback* plus *intervention* (Figure

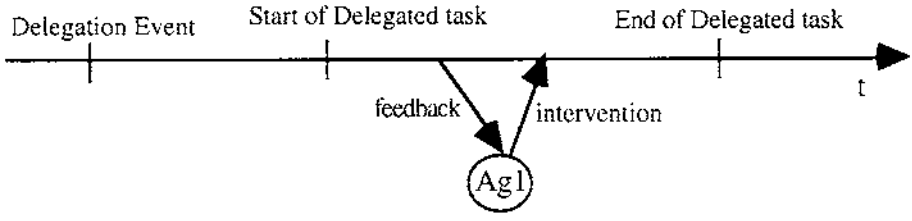

FIGURE 5. Control channels for the client's adjustment.

5).[10] 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 the intervention hold.

Feedback can be provided by observation of Ag_2's activity (inspection, surveillance, monitoring), 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* one considers five kinds of intervention:

- *stopping the task* (the delegation or the adoption process is suddenly interrupted);
- *substitution* (an intervention allocates part of (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);
- by *moving from* the purely temporal monitoring/reporting to the working phases monitoring/reporting (or vice versa).

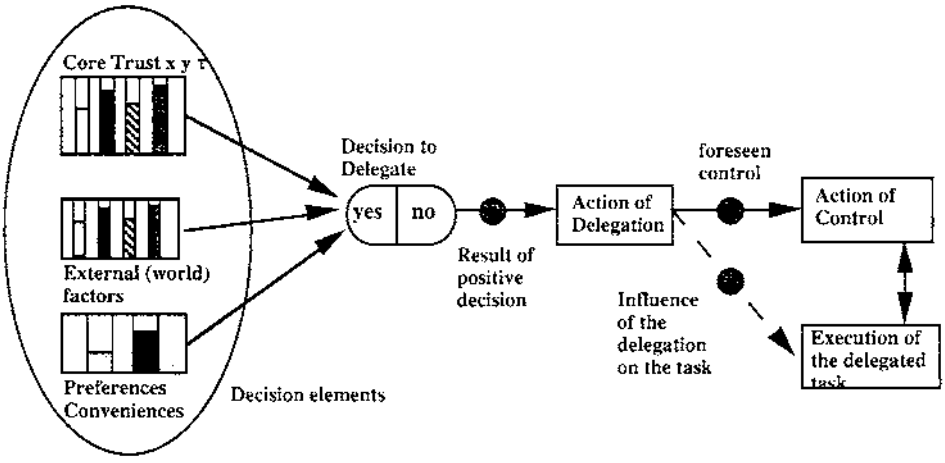

FIGURE 6. Decision, delegation, and control.

The *frequency of intervention* is also 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, the frequencies of intervention and monitoring/reporting are also 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.*

Figure 6 integrates the schema of Figure 3 with the two actions: control and execution of the task. Plans typically contain control actions of some of their actions (Castelfranchi & Falcone, 1994).

CONTROL REPLACES TRUST AND TRUST MAKES CONTROL SUPERFLUOUS?

As was said before, a perspective of duality between trust and control is very frequent and at least partially valid (Tan & Thoen, 1999). 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)[11]

This captures a very intuitive and commonsense use of the term trust (in social interaction). In fact, it is true—in this restricted 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 & Roth, 1993), or as "*functional equivalent . . . mechanisms*" (Tan & Thoen, 1999): "to reach a minimum level of confidence in cooperation, partners can use trust and control to complement each other" (Beamish, 1988).[12]

With respect to this view, there are 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 loses some important use 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.

It will be argued that:

- control is antagonistic to strict trust;
- control requires new forms of trust and builds the broad trust;
- control completes and complements it;
- control can even create and increase the strict trust.

A Strict Trust Notion (Antagonist of Control) and a Broad Notion (Including Control)

As was said, there is agreement with the idea that (at some level) trust and control are antagonistic (one eliminates the other) but complementary. This notion of trust, as defined by Mayer, is considered too restricted. It represents the notion of trust in strict sense, i.e., applied to the agent (and, in particular, to a social agent and a process or action), and strictly relative to the "internal attribution," to the internal factor. In other words, this represents 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 nontrust) that g.[13] Both forms of trust are considered. 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.

A broader notion of trust is proposed including all the 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 Figure 7).

This strict notion is similar to that 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.[14] In this model, trust in y and control of y are *antagonistic*: where there is trust there is no control, and vice versa; the larger the trust the less room for control, and vice verse. But they are also *supplementary*: one remedies to the lack of the other; they are parts of one and the same entity. What is this attitude that can either be built out of trust or out of control? It is confidence, i.e., trust again, but in a broader sense, as we formalized it.

In this view one needs these two levels and notions of trust. In this perspective, notice that control is both antagonist to (one form of trust) and constituent of (another form of) trust.

Obviously, this schema is very simplistic and just intuitive. This idea will be made more precise. However, it is immediately remarkable 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

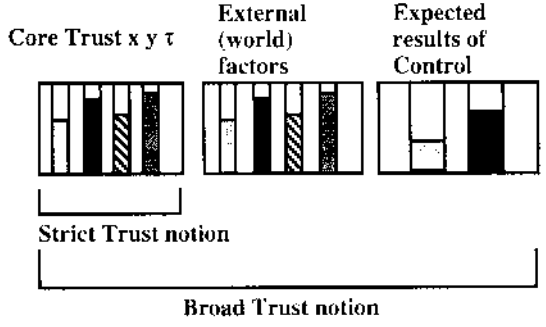

FIGURE 7. Control complements strict trust.

enough); it can also be aimed precisely at augmenting the internal trust in *y*, *y*'s trustworthiness.

Relying on Control and Bonds Requires Additional Trust

To this account of trust, one might object that the importance of trust is overstated in social actions such as contracting and organizations 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 organizations people would not rely on trust but on authorization, permission, obligations, and so forth.

In this view (Castelfranchi & Falcone, 1998a) 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, organizations) one just deals with *a more complex and specific kind of trust*. But trust is always crucial.

Control is put in place because it is one believed that the trustee will not avoid or trick monitoring, but will accept possible interventions, and be positively influenced by control. One puts a contract in place only because one believes 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 requires also that *x trusts* some authority or its own ability to monitor and sanction *y* (see Castelfranchi & Falcone, 1998a, on *three party trust*). *X* must also trust procedures and means for control (or the agent delegated to this task).

How Control Increases and Complements Trust

As one saw, control in a sense complements and surrogates trust and makes a broad trust notion (see Figure 7) sufficient for delegation and betting. How does this work? How does control precisely succeed in augmenting confidence?

One 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

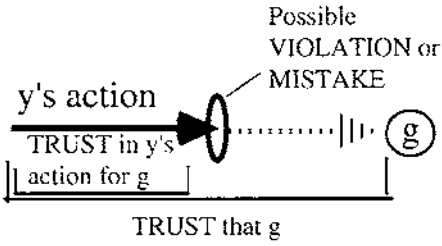

FIGURE 8. Trust in the action V's trust in the result.

somebody else. Control can be aimed at filling this gap between *y*'s intention and action and the desired result "that *g*" (Figure 8).

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 precondition of trusting *y*. Analogously, the decision to put control can affect the strict trust whose level makes 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

Two Kinds of Control[15]

Pushing or Influencing: 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 is at least one of the following beliefs:

i) If *y* is (knows to be) surveilled its performance will be better because it will either put more attention, effort, or 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 most probably will do the task.*

Since *x* believes this, by deciding to control *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 one can see in Figure 9, one of

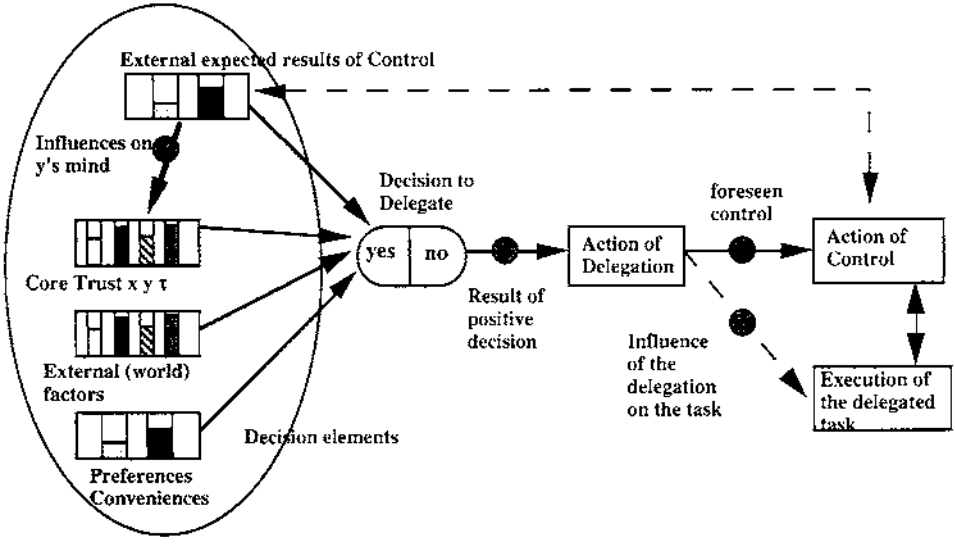

FIGURE 9. The expectation for control enters the decision of trusting.

the control results is to just change the core trust of x on y about τ. More formally one can write

$$\text{Bel}(y\ \text{Control}(x\ y\ \tau)) \supset \text{Attitudes-under-External-Control}(y\ \tau). \quad \text{(a)}$$

In other words, if y believes that x controls it about τ a set of y's attitudes will be introduced by y during its performance of τ.

In addition, if

$$\text{Bel}(x\ \text{Bel}(y\ \text{Control}(x\ y\ \tau)) \supset \text{Attitudes-under-External-Control}(y\ \tau)), \quad \text{(b)}$$

then

$$(\text{DoT}^*_{x,y,\tau} \gtreqless \text{DoT}_{x,y,\tau}),$$

where $\text{DoT}^*_{x,y,\tau}$ is the x's degree of trust of y about τ with the presence of control. For example, these additional attitudes can change y's attention, effort, care, reliability, correctness, etc. and, consequently, produce a positive, negative, or null contribution to the x's degree of trust of y about τ (depending from the expectation of x).

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.*[16] This can be just a form of "implicit communication" (to let the other see/believe that one can see him, and that one 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.

Safety Correction or Adjustment Control: Preventing Failure or Damages. This control is aimed at preventing dangers due to *y*'s violations or mistakes, and, in general, more is aimed at having the possibility of adjustment of delegation and autonomy of any type (Falcone & Castelfranchi, 2000). In other words, it is not only for repairing but for correction, through advice, new instructions and specification, changing or revoking tasks, 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. More formally,

$$\text{Control}(x\ y\ \tau) \supset \text{Ability-Intervention}(x\ y\ \tau) \qquad \text{(c)}$$

and, in general, x believes that the probability to achieve g when it is possible to intervene -Pr*(achieve(g))- is greater than without this possibility Pr(achieve(g)):

$$\text{Bel}(x\ (\text{Pr*}(\text{achieve}(g))) > (\text{Pr}(\text{achieve}(g)))).$$

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. The 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 al). 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.[17]

Moreover, we argue that an effect (and a function/aim) of the second form of control can also be 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.).

Filling the Gap Between Doing/ Action and Achieving/ Results

Let one put the problem in another perspective. As was 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 scenarios, one can say that the first kind of control is pointing to, is impinging on the first step (trust in y), and is aimed at increasing 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 counting on y, but x counts on a multiagent possible plan that includes possible actions of it.

As one can see from formula (a) the important thing is that y believes that the control holds, and not if it 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 (Figure 10).

While there is control for correction/adjustment, the achievement of g is committed to y's action plus x's possible action (intervention), x bets on this combination (Figure 11).

Very similar complementing or remedying roles are guarantees, protections, and assurance. One does not trust the action enough, and one puts protections in place to be sure about the desired results. For example, one does not trust driving a motorcycle without a crash helmet, but one trusts doing so with it.

FIGURE 10. The gap between action and expected results.

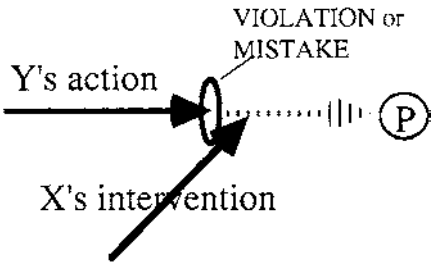

FIGURE 11. Intervention in the gap.

The Dynamics

It is important to underline 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 carefulness, 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 surveils y) y will be more committed, willing, and reliable; i.e., the strength of x's trust beliefs in y and thus x's degree of trust in y are improved.

This is very interesting social (moral and pedagogical) strategy. In fact, it is in opposition to another well-known strategy aimed at increasing y's trustworthiness—i.e., "trust creates trust!"[18] In fact, precisely the reduction/renouncement to control is a strategy of "responsibility" of y, aimed at making it more reliable, more committed.

Those strategies are in conflict with each other. When and why do we choose to make y more reliable and trustworthy through responsibility (renouncement 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 cyberpartners) more reliable and trustworthy through responsibility or through surveillance? There will not be 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, those subtle interaction problems will be relevant for sure for computer-mediated human interaction and collaboration.

Control Kills Trust

Control can be bad and self-defeating, in several ways:

- There might be misunderstandings, mistakes, and incompetence and wrong intervention by the controller ("who does control controllers?") (in this case $\Pr^*(\text{achieve}(g)) < \Pr(\text{achieve}(g))$.
- Control might have the opposite effect than function (A), i.e., instead of improving performance, it might make performance worse. For example, by producing anxiety in the trustee or by making him waste time and concentration in preparing or sending feedbacks (case in which $\text{DoT}^* < \text{DoT}$).

- It can produce a breakdown of willingness. Instead of reinforcing commitment and willingness, control can disturb it because of reaction or rebellion, or because of delegation conflicts (Castelfranchi & Falcone, 1997) and need for autonomy, or because of the fact that distrust creates distrust (also in this case DoT* < DoT).

Here, one cares mainly of the bad effect of control on trust; thus let us see these dynamics. As trust virtuously creates trust, analogously the trust of y in x, which can be very relevant for his motivation (for example, in case of exchange and collaboration), can decrease because x exhibits not so much trust in y (by controlling y).

– X is too diffident. Does this mean that x is malicious and machiavellic? Since x suspects so much about the others would be ready for deception? Thus, if x distrusts y, y can become diffident about x.
– Otherwise, x is too rigid, not the ideal person to work with.[19]
– Finally, if the agents rely on control, authority, norms, they relax the moral, personal, or affective bonds, i.e., one of the strongest bases for interpersonal trust. Increasing control procedures in organizations and community can destroy trust among the agents, and then make cooperation, market, organization very bad or impossible, since a share of risk acceptance and trust is unavoidable and vital.

In sum, as for the dynamics of such a relation, it was explained how

- x's control of y denounces and derives from a lack of x's trust *in* y;
- x's control of y can increase x's trust *in* y;
- x's control of y increases x's trust in deciding to delegate to y (his global trust);
- control of y by x can both increase and decrease y's trust in x; in case control decreases y's trust in x, this should also affect x's trust in y (thus this effect is the opposite of the second);
- x's control of y improves y's performance or makes it worse;
- x's control of y improves y's willingness or makes it demotivated.

CONCLUSIONS

Does control reduce or increase trust? As one 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 it was claimed that control is antagonistic of this strict form of trust, but that it also completes and complements it for arriving at a global trust. In other words, putting control and guarantees in trust-building. It produces a sufficient trust, when trust in *y*'s autonomous willingness and competence would not be enough. It has also been argued that control requires new forms of trust; trust in the control itself or in the controller, trust in *y* for being monitored and controlled, trust in possible authorities, etc.

Finally, it has been 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.

Two kinds of control were also analyzed, characterized by two different functions: *pushing or influencing control* aimed at preventing violations or mistakes, versus *safety, correction, or adjustment control* aimed at preventing failure or damages after a violation or mistake.

NOTES

1. There may be delegation without trust: these are exceptional cases in which either the delegating agent is not free (coercive delegation) or he has no information and alternative to delegating, so that he must just make a trial (blind delegation).
2. The trust that *g* does not necessarily require the trust in *y*. One must ignore which are the causal factors producing or maintaining *g* true in the world; nevertheless, one may desire, expect, and trust that *g* happens or continues. The trust that *g*, per se, is just a—more or less supported—subjectively certain positive expectation (belief conform to desire) about *g*.
3. We use the classical modal logic operators and the constructs of dynamic logic. In particular, the belief about practical opportunity borrows from dynamic logic the construct $\langle \mathrm{Do}_i(\alpha) \rangle g$ that means that agent *i* has the opportunity to perform action α in such a way that *g* will result from this performance.
4. Moreover, there might not only be the frustration of *g*, the missed gain, but there might be additional damages as effect of failure, negative side effects: the risks in case of failure are not the simple counterpart of gains in case of success.
5. Beliefs 2a and 2b imply some beliefs about *y*'s motives: intention is due to these motives and persistence is due to preferences between motives.
6. To be true, we should also consider the reciprocal influence between external and internal factors. When *x* trusts the internal powers of *y*, it also trusts its abilities to create positive opportunities for success, to perceive and react to the external problems. Vice versa, when *x* trusts the environment opportunities, this valuation could change the trust about *y* (*x* could think that *y* is not able to react to specific external problems).
7. In all forms of adoption-based trust, beliefs about adoptivity and motives for adoption are particularly crucial.
8. $\mathrm{U}(\mathrm{Ag}_1)_t$, is the Ag_1's utility function at the time *t*, and, specifically: $U(\mathrm{Ag}_1)_{p+,t}$, the utility of the Ag_1's success performance; $U(\mathrm{Ag}_1)_{p-,t}$, the utility of the Ag_1's failure performance; $U(\mathrm{Ag}_1)_{\mathrm{di}+,t}$ the utility of a successful delegation to the agent *i* (the utility due to the success of the delegated action); $U(\mathrm{Ag}_1)_{\mathrm{di}-,t}$ the utility of a failure delegation to the agent *i* (the damage due to the failure of the delegated action); $U(\mathrm{Ag}_1)_{0,t}$ the utility to do nothing.
9. More precisely, we have: $U(\mathrm{Ag}_1)_{p+,t}$ = Value(*g*) + Cost [Performance(Ag_1 *t*)], $U(\mathrm{Ag}_1)_{p-,t}$ = Cost [Performance(Ag_1 *t*)] + Additional Damage for failure, $U(\mathrm{Ag}_1)_{\mathrm{di}+,t}$ = Value(*g*) + Cost [Dele-

gation(Ag_1 Ag_i τ t)], $U(Ag_1)_{di-,t}$ = Cost [Delegation (Ag_1 Ag_i τ t)] + Additional Damage for failure, where it is supposed that it is possible to attribute a quantitative value (importance) to the goals and where the costs of the actions (delegation and performance) are supposed to be negative.

10. *Control activity* will be called the combination of two more specific activities: monitoring and intervention.
11. This is a remarkable definition. The authors' analytical account is rather close to it: "a particular action important to the trustor" means that the trustor has some goal and is relying on such an action of the trustee for such a goal (delegation); trust is a "willingness," a decision to bet on somebody, to take some risk, then to be "vulnerable," but is based on expectations about the willingness of the other party. One just considers "trust" as ambiguous: able to designate on the one side the decision and the action of relying; on the other side, the mental attitude toward that party that is presupposed by such a decision, i.e., those "expectations." Moreover, one takes into account not only the willingness but also the competence (and even the external opportunities). Finally, one does not put the restriction of "irrespective to control," because the definition is more general and goes beyond strict social/personal trust in the other. Here, one clarifies precisely this point by proposing a strict and broad notion of trust.
12. Of course, as Tan and Thoen (1999) noticed, control can be put in place by default, not because of a specific evaluation of a specific partner, but because of a generalized rule of prudence or for lack of information. (See later, about the level of trust as insufficient either for uncertainty or for low evaluation.)
13. Somebody, call this broader trust "confidence." But, in fact, they seem quite synonymous: there is confidence *in y* and confidence *that p*.
14. Control—especially in collaboration—cannot be completely eliminated and lost, and delegation and autonomy cannot be complete. This is not only for reasons of confidence and trust, but for reasons of distribution of goals, knowledge, competence, and for an effective collaboration. The trustor usually has to know at least whether and when the goal has been realized or not.
15. There is a third form of control (or better of monitoring) merely aimed at *y* evaluation. If this mere monitoring (possibly hidden to *y*) is for a future adjustment off-line (for changing or revocating the delegation next time), this form of control becomes of *B* kind: control for adjustment, for correction.
16. It is also necessary that *y* cares about *x*'s evaluation. Otherwise, this control has no efficacy. A bad evaluation as some sort of "sanction," however, is not an "intervention"—except if *x* can communicate it to *y* during its work, since it does not interrupt or affect *y*'s activity.
17. Different kinds of delegation allow for specific functions of this control. There will be neither compensation nor sanctions in weak delegation (no agreement at all), while there will be intervention for remedy.
18. Trust creates trust in several senses and ways. The decision to trust *y* can increase *x's trust* in a way, via several mechanisms: cognitive dissonance; because *x* believes that *y* will be responsible; because *x* believes that *y* will feel more self-confident; because *x* believes that *y* will trust *x* and then bring more goodwill. The decision to trust *y* can increase *y's trust* in a way, via several mechanisms: *y* has power over *x* that makes himself vulnerable and dependent; *y* feels that if *x* is not diffident he is probably not malicious; *y* perceives a positive social attitude by *x* and this elicits his goodwill, etc. However, this is not the right place for developing this theory.
19. Control could also increase *y*'s trust in *x*, as a careful person, or a good master and boss, etc.

REFERENCES

Beamish, P. 1998. *Multinational joint ventures in developing countries*. London: Routledge.

Bons, R., F. Dignum, R. Lee, and Y. H. Tan. 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 Societies,"* Minneapolis, MN, May 9, 21–34.

Castelfranchi, C. 1998. Modeling social action for AI agents. *Artificial Intelligence* 103:157–182.

Castelfranchi, C. 2000. *Formalizing the informal*? Invited talk International Workshop on Deontic Logic (DEON 2000) Toulouse.

Castelfranchi, C., and R. Falcone. 1994. Towards a theory of single-agent into multi-agent plan transformation. *Third Pacific Rim International Conference on Artificial Intelligence* (*PRICAI94*), Beijing, China, 16–18 August, 31–37.

Castelfranchi, C., and R. Falcone. 1997. Delegation conflicts. In *Multi-agent rationality*, eds. M. Boman and W. Van de Velde. *Lecture Notes in Artificial Intelligence*, 1237:234–254. New York: Springer-Verlag

Castelfranchi, C., and R. Falcone. 1998a. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings of the International Conference on Multi-Agent Systems* (*ICMAS'98*), Paris, July, 72–79.

Castelfranchi, C., and R. Falcone, 1998b. Towards a theory of delegation for agent-based systems. In *Robotics and autonomous and systems*, Elsevier Editor 24(3–4):141–157.

Castelfranchi, C., and R. Falcone. 2000. Social trust: A cognitive approach. In *Deception, fraud and trust in virtual societies*, eds. C. Castelfranchi and Yao-Hua Tan. Norwell, MA: Kluwer Academic Publisher (in press).

Castelfranchi, C., and Y.-H. Tan. 2000. Introduction. In *Deception, fraud and trust in virtual societies* eds. C. Castelfranchi and Y.-H. Tan. Norwell, MA: Kluwer Academic Publisher (in press).

Dasgupta, P., 1990. Trust as a commodity. In *Trust*, ed. D. Gambetta, Chapter 4, 49–72. Oxford: Basil Blackwell.

Deutsch, M. 1973. *The resolution of conflict*. New Haven, CT: Yale University Press.

Falcone, R., and C. Castelfranchi. 2000. Levels of delegation and levels of adoption as the basis for adjustable autonomy. *Lecture Notes in Artificial Intelligence* 1792:285–296.

Hendler, J. 1999. Is there an intelligent 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* 3:223–250.

Luhmann, N. 1990. Familiarity, confidence, trust: Problems and alternatives. In *Trust*, ed. D. Gambetta, Chapter 6, 94–107. Oxford: Basil Blackwell.

Marsh, S. 1994. *Formalising trust as a computational concept*, Ph.D. thesis, Department of Computing Science, University of Stirling, Scotland.

Mayer, R. C., J. H. Davis, and F. D. Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review* 20(3):709–734.

Meyer, J. J., and W. van der Hoek. 1992. A modal logic for nonmonotonic reasoning. In *Non-monotonic reasoning and partial semantics*, eds. W. van der Hoek, J. J. Ch. Meyer, Y. H. Tan, and C. Witteveen, 37–77. Chichester: Ellis Horwood.

Nissenbaum, H. 1999. Can trust be secured online? A theoretical perspective. http://www.univ.trieste.it/~dipfilo/etica_e_politica/1992_2/nissenbaum.html

Raub, W., and J. Weesie. 1990. Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology* 96:626–654.

Rea, T. 2000. Engendering trust in electronic environments—Roles for a trusted third party. In *Deception, fraud and trust in virtual societies*, eds. C. Castelfranchi and Y.-H. Tan. Norwell, MA: Kluwer Academic Publisher (in press).

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

Sitkin, S. B., and N. L. Roth. 1993. Explaining the limited effectiveness of legalistic "remedies" for trust/distrust. *Organization Science* 4:367–392.

Snijders, C., and G. Keren. August 1996. Determinants of trust. In *Proceedings of the Workshop in Honor of Amnon Rapoport*, University of North Carolina at Chapel Hill, 6–7.

Tan, Y.-H., and W. Thoen. 1999. Towards a generic model of trust for electronic commerce. *Autonomous Agents '99 Workshop on "Deception, Fraud and Trust in Agent Societes,"* Seattle, WA, May 1, 117–126.

van Linder, B. 1996. *Modal logics for rational agents*, Ph.D. thesis, Department of Computing Science, University of Utrecht.