7

Trust, Control and Autonomy: A Dialectic Relationship

In this chapter we are going to analyze the relationships between trust, control and autonomy: in particular, we are interested in showing how *trust and control are strictly interwined*, how their relationships are dynamic and influence the autonomy of the involved agents. We will also analyze the concept of 'adjustability' of both autonomy and delegation, and how it is dependent, elicited and guided from the previous notions of control and trust and from their interactions.

7.1 Trust and Control: A Complex Relationship

The relationship between trust and control is quite relevant both for the very notion of trust and for modelling and implementing trust-control relationships among autonomous systems; but it is not trivial at all.

On the one hand, it is true that where/when there is monitoring and control there is no trust (or at least there is less trust than without control), and vice versa: when/where there is a deliberate absence of control, there is trust (or at least there is more trust than in the case in which it has been necessary to insert control). However, this refers to a restricted notion of trust: i.e., what we call 'trust in Y', which is just a part, a component of the global trust needed because of relying on the action of another agent. We claim that *control is antagonistic of this strict form of trust* (internal trust, see Chapter 2); but also that it can complete and complement strict trust (in Y) for arriving at a *global* trust. In other words, putting control and guarantees is an important step towards 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 (or third parties; Section 7.1.5); and so on.

Finally, we show that paradoxically control might not be antagonistic of strict trust in Y, but it could even create trust, increase it by making Y more willing or more effective.

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

We will show how, depending on the circumstances, control makes *Y* more reliable or less reliable; control can either decrease or increase *Y*'s trustworthiness and the internal trust. We will also analyze two kinds of control, 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 good theory of control and of their reciprocal interactions.

7.1.1 To Trust or to Control? Two Opposite Notions

The relation between trust and control is very important and perhaps even definitory; however it is everything but obvious and linear. On the one hand, some definitions delimit trust precisely thanks to being its opposite. But it is also true that monitoring and guarantees 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?[1]

On the other hand, it appears that the 'alternative' between control and trust is one of the main *tradeoffs* in several domains of information technology and computer science, from Human Computer Interaction to Multi-Agent Systems, Electronic Commerce, Virtual Organisations, and so on, precisely as in human social interaction.

Consider, for example, the problem of mediating between two such 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, authorization, cryptography, inspection, control, etc. (Castelfranchi, 2000) which can be in fact self-defeating for improving Electonic Commerce, 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.

7.1.2 What Control Is

'Control' is a (meta) action[2]:

[1] *'Do you trust this system/company/aircraft/drug . . . !?' 'Yes I do! There are so many controls and safety measures . . !'*

[2] We will call *control activity* the combination of two more specific activities: monitoring and intervention.

(a) aimed at ascertaining whether another action has been successfully executed or if a given state of the world has been realized or maintained (*monitoring, feedback*);
(b) aimed at dealing with the possible deviations and unforeseen events in order to positively cope with them and adjusting the process (*intervention*).

When the trustor is delegating (see Section 2.6) a given object-action, what about its control activity? Considering, for the sake of simplicity, that the control action is executed by a single agent, if *Delegates*(Ag_1 Ag_2 τ) there are at least four possibilities:

i) Ag_1 delegates the control to Ag_2: the trustor delegates both the task and the control on the task realization to the trustee;
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 one of these possibilities could be either about (a) (monitoring, feedback) or (b) (intervention), and could be either 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 previous interactions between the trustor and trustee, etc.).

To understand the origin and the 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 the position 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 (see Figure 7.1).

Otherwise no adjustment is possible. Obviously, the feedback useful for a run-time adjustment must be provided in time for the intervention. In general, the feedback activity is the

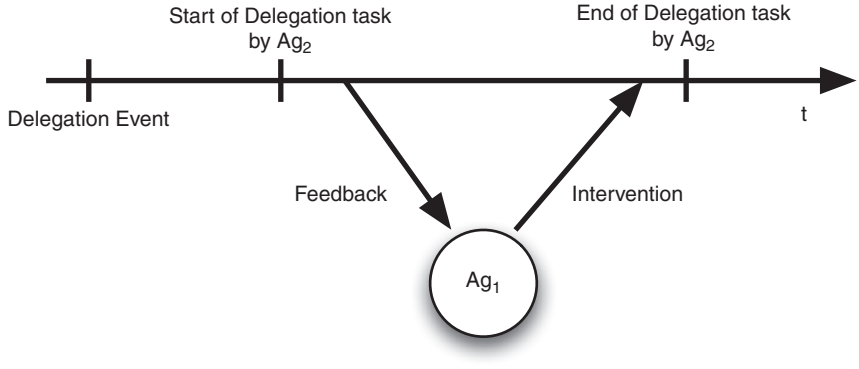

Figure 7.1 Control channels for the client's adjustment. (Reproduced by Permission of © 2001 IEEE)

precondition for an intervention; however it is also possible that either only the feedback or only the intervention will hold.[3]

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 possible kinds:

i) *stopping the task* (the delegation or the adoption process is suddenly interrupted by the trustor);
ii) *substitution* (an intervention by the trustor allocates part of the (or the whole) task either to the trustor themselves or to a third agent);
iii) *correction of delegation* (after the intervention by the trustor, the task is partially or totally changed: the intervention transforms/changes the delegated task without any change of task allocation to other agents);
iv) *specification or abstraction of delegation* (after the intervention by the trustor, the task is more or less constrained; this is a specific case of the previous kind (*correction of delegation*));
v) *repairing of delegation* (the intervention by the trustor leaves the task activity unchanged but it introduces new actions (that have to be realized by either the trustee, or the trustor or some other agent) necessary to achieve the goal(s) of the task).

Imagine that Ag_1 and Ag_2 have decided to prepare a dinner at home, and Ag_1 delegated the task of cooking 'pasta with pesto' to Ag_2 while Ag_1 is preparing two tomato eggs; we have:

- case (i) when for example suddenly Ag_1 stops this delegation to Ag_2 (maybe Ag_1 is no longer hungry, she feels unwell, someone else has brought pizza to their house, and so on);
- case (ii) when for example Ag_1 decides to prepare the pesto herself (maybe Ag_2 is not able to find the ingredients, he is too slow, he is not able to mix the different parts correctly);
- case (iii) when for example Ag_1 sees that the basilico is finished and suggests to Ag_2 that they (or she prepares herself) the 'aglio e olio' as a sauce for spaghetti;
- case (iv) when for example Ag_1 seeing that Ag_2 is not completely happy about the spaghetti with pesto says to him to prepare spaghetti with the sauce he prefers;
- case (v) when for example Ag_1 seeing that the pesto sauce prepared by Ag_2 is not enough for two people, prepares an additional quantity of pesto.

Each of these interventions could be realized through either a *communication act* or a *direct contribution* to the task by the trustor.

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 into phases and the monitoring or the reporting is connected with them).

[3] Sometimes we want to monitor the delegated action or its result not in time and in order for intervention. But just for the future; for confirming or correcting out trust in Y (see later).

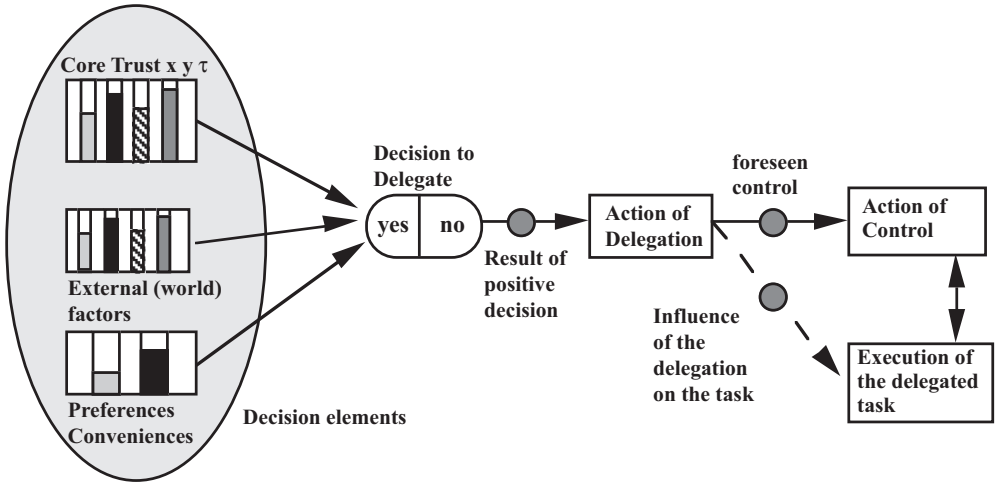

Figure 7.2 Decision, delegation and control

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*.

Figure 7.2 shows how the control action impacts on the execution of the task after the trustor's delegation to the trustee. Plans typically contain control actions of some of their actions (Castelfranchi and Falcone, 1994).

7.1.3 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 (Tan and Thoen, 1999). Consider for example the definition of (Mayer *et al.*, 1995):[4]

> 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

[4] About our more analytic considerations on this definition see Chapter 1 in this book.

This captures a very intuitive and common sense use of the term trust (in social interaction). In fact, it is true – in this limited sense – that if you control me 'you don't trust me!'; and it is true that if you do not trust me enough (to count 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' (Tan and Thoen, 1999): 'to reach a minimum level of confidence in cooperation, partners can use trust and control to complement each other' (Beamish, 1988).[5]

We have some problems with respect to this view:

- on the one hand, 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 hand, there is something reductive and misleading in such a position:
  - it reduces trust to a strict notion and loses 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 and his trustworthiness.

We will argue that:

- firstly, control is *antagonistic* to strict trust;
- secondly, it requires new forms of trust including *broad trust* to be built;
- thirdly, it *completes and complements* it;
- finally, it can even create, *increase* the strict/internal trust.

As the reader can see it is quite a complex relationship.

7.1.4 Trust Notions: Strict (Antagonist of Control) and Broad (Including Control)

As 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 a 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, it represents the 'trust *in Y*' (as for action α and goal g) (see Section 2.6.1). But this trust – when 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.[6] 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 do the action α appropriately

[5] 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, on the level of trust as insufficient either for uncertainty or for low evaluation).

[6] Somebody, call this broader trust 'confidence'. But in fact they seem quite synonymous: there is confidence *in Y* and confidence *that g*.

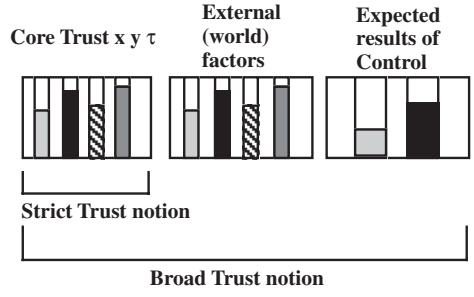

Figure 7.3 Control complements strict trust

(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; including actions of other agents, and including possible control activity on *Y*) 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 7.3).

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 hand they are also two complementary parts, as for the broad/global trust: control supplements trust.[7]

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 versa; 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.[8]

In our view we need these two levels and notions of trust. With this in mind, notice that control is both *antagonist* to (one form of trust: the *strict* one) and *consituent* to (another form of trust: the *broader* one). Obviously, this schema is very simplistic and just intuitive. We will make this idea more precise. However, let us note 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 also be aimed precisely at augmenting the internal trust in *Y*, *Y*'s trustworthiness.

[7] Control – especially in collaboration – cannot be completely eliminated and lost, and delegation and autonomy cannot be complete. This, not only for reasons of confidence and trust, but for reasons of distribution of goals, of knowledge, of competence, and for an effective collaboration. The trustor usually has at least to know whether and when the goal has been realized or not (Castelfranchi and Falcone, 1994).

[8] This holds for a fully delegated task. It is clear that for coordination between X and Y in a multi-agent plan, X *has to* monitor Y (and vice versa) even if she trusts him a lot.

7.1.5 *Relying on Control and Bonds Requires Additional Trust: Three Party Trust*

To our account of trust one might object that we overstate the importance of trust 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 our view (Castelfranchi and Falcone, 1998) 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) we are just dealing with *a more complex and specific kind of trust.* But trust is always crucial.

As Emile Durkheim claims 'A contract is not sufficient by itself, but is only possible because of the regulation of contracts, which is of social origin' ((Durkheim, 1893) p. 162), and this social background includes trust, social conventions and trust in them, and in people respecting them, the authorities, the laws, the contracts (see Chapter 9).

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 requires also that *X trusts* some authority or its own ability to monitor and to sanction *Y*, see (Castelfranchi and Falcone, 1998). *X* must also trust procedures and means for control (or the agent delegated to this task).

To be absolutely clear, we consider this level of trust as a *three party relationship*: it is a relation between the client *X*, the contractor *Y* and the authority *A*. And there are three trust

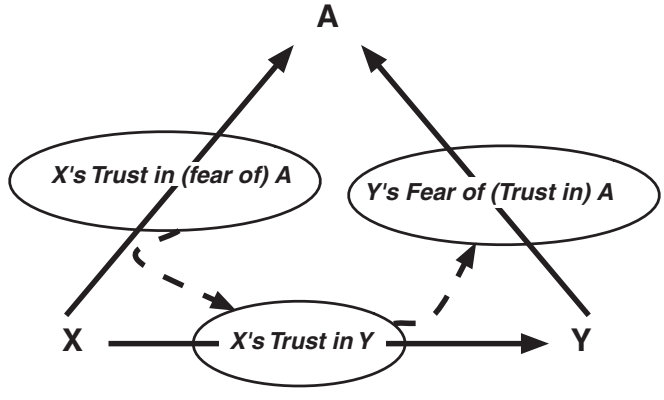

Figure 7.4 Three party relationships among Trustor, Trustee and Authority

sub-relations in it see Figure 7.4):

- *X trusts Y* by believing that *Y* will do what is promised because of his honesty or because of his respect/fear toward *A*;
- *X trusts A* and its ability to control, to punish etc. and relies on *A* for this;
- *Y trusts* A (both when *Y* is the client and when he is the contractor) the same beliefs being the bases of his respect/fear toward *A* (that is: trusting a threatening agent!).

In other words, *X* relies on a form of *paradoxical* trust of *Y* in *A*: *X* believes that *Y* believes that *A* is able to control, to punish, etc. Notice that *Y*'s beliefs about *A* are precisely *Y*'s trust in the authority when he is the client. When *Y* is the contractor the same beliefs are the bases of his respect/fear toward *A*.

We can also say that, in addition to the *X*'s trust in *Y* based on the internal (*Y*'s competence and willingness) and contextual-environmental reasons believed by *X*), there is a part of *X*'s trust in *Y* based on the fact that the other two relationships are true (and believed by the agents) and are the relationships of *X* and *Y* with the authority *A*.

In sum, in contracts and organizations it is true that *personal trust* in *Y* may not be enough, but what we put in place is a higher level of trust which is our trust in the authority but also our trust in *Y* as for acknowledging, worrying about and respecting the authority. Without this trust in *Y* the contract would be useless. This is even more obvious if we think of possible alternative partners in contracts: how to choose among different contractors with the same conditions? Precisely on the basis of our degree of trust in each of them (both, trust about their competence, but also trust about their reliability, their respecting the contract).

As we have already said, these more complex kinds of trust are just richer specifications of the reasons for *'Y*'s doing what we expect: reasons for *Y* 's predictability which is based on his willingness; and reasons for his willingness (he will do α, either because of his selfish interest, or because of his friendliness, or because of his honesty, or because of his fear of punishment, or because of his institutional and normative respect: several different bases of trust).

More formally (simplifying with respect to the external conditions and concentrating on the core trust, as shown in Figures 2.11, 2.12, and 2.13 in Chapter 2):

X's mental state in *Trust(X, Y, τ)* is essentially constituted by:

$$Bel_X Can_Y(\alpha, p) \wedge Bel_X WillDo_Y(\alpha, p) \tag{7.1}$$

X's mental state in *Trust(X, A, τ')* is essentially constituted by:

$$Bel_X Can_A(\alpha', p') \wedge Bel_X WillDo_A(\alpha', p') \tag{7.2}$$

Y's mental state in *Trust(Y, A, τ')* is essentially constituted by:

$$Bel_Y Can_A(\alpha', p') \wedge Bel_Y WillDo_A(\alpha', p') \tag{7.3}$$

Where τ is the task that *Y* must perform for *X;* τ' the task that *A* must perform for *X* towards *Y*, i.e. check, supervision, guarantee, punishment, etc.

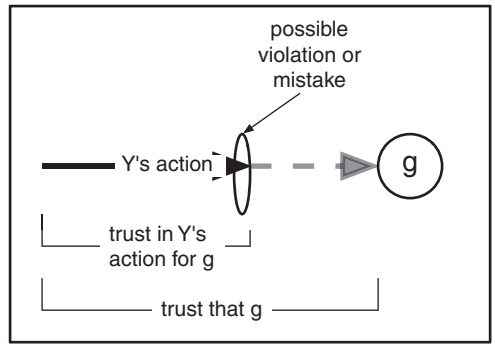

Figure 7.5 Trust in *Y*'s action versus Trust in the final Result (achieving the goal)

More precisely and importantly in *X's* mind there is a belief and a goal (thus an expectation) about this trust of *Y* in *A:*

$$Bel_X(Trust(Y, A, \tau')) \wedge Goal_X(Trust(Y, A, \tau'))^9 \tag{7.4}$$

And this expectation gives an important contribution to *X*'s trust in the contractor.

X trusts *Y* by believing that *Y* will do what is promised because of his honesty or because of his respect/fear toward *A*. In other words, *X* relies on a form of paradoxical trust of *Y* in *A: X* believes that *Y* believes that *A* is able to control, to punish, etc. Of course, normally a contract is bilateral and symmetric, thus the point of view of *Y's* should be added, and his trust in *X* and in *A* as for monitoring *X*.

7.1.6 How Control Increases and Complements Trust

As we saw, in some sense control complements and surrogates trust and makes broad trust notions (see Figure 7.3) sufficient for delegation and betting. How does this work? How does control precisely succeed in augmenting confidence?

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' (Figure 7.5).

However, does control only augment the broad trust? Not necessarily: 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 an ante-hoc and static datum (either sufficient or insufficient for delegation before the decision to delegate). It is a dynamic entity (see Chapter 6 in this

[9] As we said, here the use of the predicate 'Trust' is a bit inappropriate and misleading in a way. Actually, this is not trust but 'fear' of *A*. We want here just to stress that the basic cognitive constitutents are the same: an *evaluation* and an *expectation* towards *A*. It just depends on the Goal implied by the expectation if this is fear or trust: it is trust when *Y* is the trustor relying on *X* (and *A*); it is fear when *Y* is the trustee.

volume). For example there are effects, feedbacks of the decision to delegate on its own precondition of trusting *Y*. Analogously the decision to put control can affect *Y*'s trustwortiness and thus the strict-trust whose level makes control necessary! Thus the schema: 'trust plus control' is rather simplistic, static, a-dialectic; since the presence of control can modify and affect the other parameters. As we wrote, there are indeed two kinds and functions of control: let us analyze these more deeply.

7.1.7 Two Kinds of Control[10]

(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 by increasing *Y*'s (perceived) trustworthiness. 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. Behind this kind of surveillance there is at least one of the following beliefs:

i) if *Y* is (knows to be) surveilled his performance will be better because either he will put more attention, or more effort, or more care, etc. in the execution of the delegated task; in other words, *he will do the task better* (there will be an influence on *Y*'s ability); or
ii) if *Y* is (knows to be) surveilled he will be more reliable, more faithful to his commitment, less prone to violation; in other words, *he probably will have a stronger intention to do the task* (there will be an influence on *Y*'s willingness).

Since *X* believes this, by deciding to control *Y* (and letting *Y* know about this) she increases her own evaluation/expectation (i.e., her trust) of *Y*'s willingness, persistence, and quality of work. As we can see in Figure 7.6, one of the control results is just to change the core trust of *X* on *Y* about τ.

More formally we can write:

$$Bel_Y(Control(X\ Y\ \tau)) \supset \Delta Trustworthiness(Y\ \tau) \tag{7.5a}$$

where *Control(X Y* τ*)* measures the level of control by *X* on *Y* about the task τ. While Δ*Trustworthiness(Y* τ*)* measures the corresponding *Y*'s variation of trustworthiness. In other words, if *Y* believes that *X* controls him about τ a set of *Y*'s attitudes will be introduced (consciously or unconsciously) by *Y* himself during his performance of τ.

In addition, if:

$$Bel_X(Bel_Y(Control(X\ Y\ \tau)) \supset \Delta Trustworthiness(Y\ \tau)) \tag{7.5b}$$

then $DoT^*_{XY\tau}$ (the *X*'s degree of trust in *Y* about τ including the knowledge of the control presence) might be different from the $DoT_{XY\tau}$ (the one without (*X*'s believed) control).

In other words, these additional attitudes can change *Y*'s attention, effort, care, reliability, correctness, etc. and consequently produce a positive, negative, but also null contribution to *X*'s degree of trust in *Y* about τ (depending from the expectation of *X*).

[10] As we said, there is a third form of control (or better of monitoring) merely aimed at *Y's* 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 the second (B) class: control for adjustment, for correction.

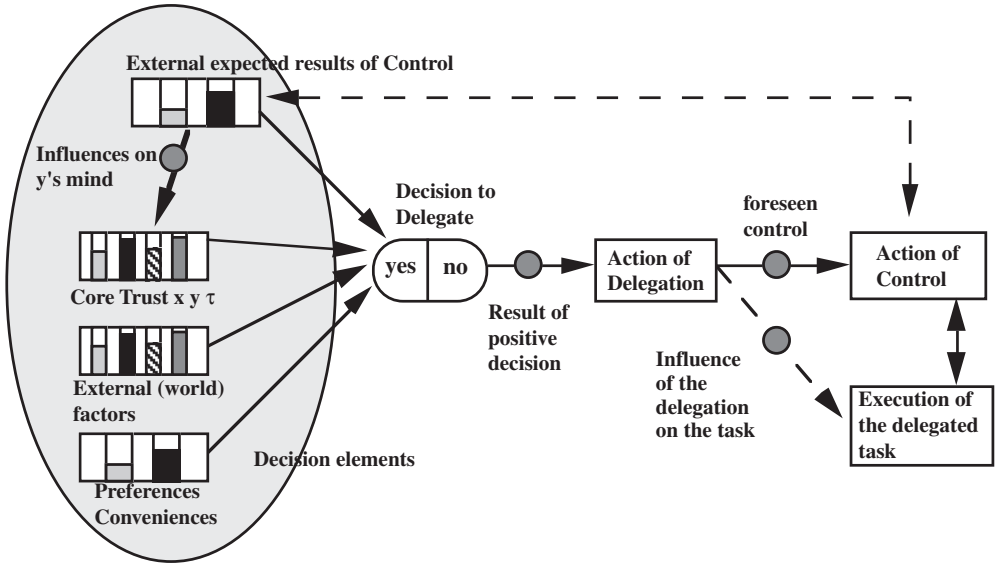

Figure 7.6 The expectation about control results influences the decision of trusting

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*.[11] 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 over this is useful (*'Don't forget that I see you!'*). Thus, some form of *intervention* can also be necessary: and as a consequence there would be 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 is aimed in general at the possibility of having adjustment of delegation and autonomy of any type ((Falcone and Castelfranchi, 2001), (Castelfranchi and Falcone, 2000b)). In other words, it is not only for repairing but for correction, through advice, new instructions and specifications, changing or revoking tasks, direct reparation, recovery, or help, etc.

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

In general *X* believes that the probability of achieving *g* when it is possible to intervene – *Pr*(achieve(g))* – is greater than without this possibility: *Pr(achieve(g))*:

$$Bel_X(Pr^*(achieve(g)) > Pr(achieve(g))) \tag{7.6}$$

[11] It is also necessary that *Y* cares about *X*'s evaluation. Otherwise this control has no efficacy. A bad evaluation is some sort of 'sanction', however it 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.

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 it is more general.

The first form/finality of control (*kindA*) prevents not only violations (in case of norms, commitments, or contracts) but also missed execution or mistakes (also in weak delegation where there are no obligations at all).

The second form/finality (*kindB*) is not only for sanctions or claims, but for timely intervening and preventing additional damages, or remedying and correcting. 'Detection' is just a means; the real aim is intervention for safety, enforcement or compensation.[12] Moreover, 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 he worries about this (sanctions, reputation, lack of autonomy, etc.).

7.1.8 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 these two components or two logical steps scenario, we 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 mainly pointing to the second step and is aimed at increasing it, by confirming the achievement of g also in case of (partial) defaillance of Y.

In this way the control (monitoring plus intervention) complements the trust in Y which would be insufficient for achieving g, and for delegating; this additional assurance (the possibility to correct work in progress of Y's activity) makes X possible to delegate to Y the goal g. In fact, in this case X is not only counting on Y, but X counts on a potential multi-agent plan that includes her own possible actions.

As we can see from the formula (7.5a) in Section 7.1.7 the important thing is that *Y believes* that the control holds, and not if it really holds.[13] For example, X could not trust Y enough and communicate to him the control: this event modifies Y's mind and X's judgment about *Y's* trustworthiness. Thus, in trust-reliance, without the possibility of intervention for correction and adjustment, there is only one possible way to achieve g, and one activity (Y's activity) on which X bets (Figure 7.7).

Meanwhile, if 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 7.8).

A very similar complementing or remedying role are guarantees, protections and assurance. I do not trust the action enough, and I put protections in place to be sure about the desired

[12] Different kinds of delegation (weak, mild, strong: see Section 2.9.1) 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.

[13] This is the actual power of the Gods.

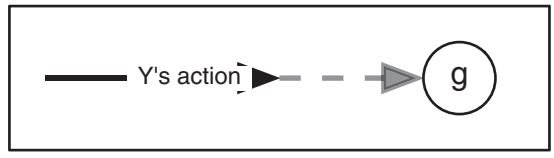

Figure 7.7 The gap between action and expected results

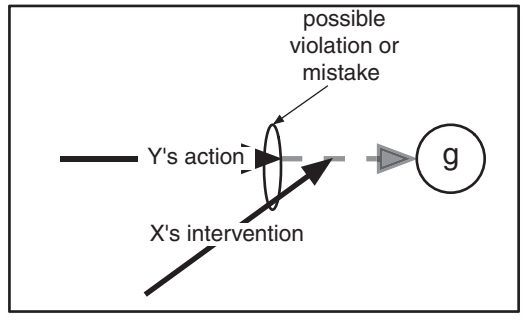

Figure 7.8 Intervention in the gap

results. For example, I do not trust driving a motorcycle without a crash-helmet, but I trust doing so with it.

7.1.9 The Dynamics

It is important to reinforce the idea 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* watches over *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 a 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' (see Chapter 6).[14]

In fact, the reduction/renouncing of control is a strategy of 'responsabilization' of *Y*, aimed at making it more reliable, more committed. Those stategies are in conflict with each other. When and why do we choose to make *Y* more reliable and trustworty through responsabilization

[14] Resuming and simplifying trust creates trust in several senses and ways. The decision to trust *Y* can increase *X*'s trust in *Y*, via several mechasnisms: cognitive dissonance; because *X* believes that *Y* will be responsabilized; because *X* believes that *Y* will feel more self-confident; because *X* believes that *Y* will trust *X* and then be more willing to do good. The decision to trust *Y* can increase *Y*'s trust in *Y*, via several mechanisms: *Y* has power over *X* that makes himself vulnerable and dependent; *Y* feels that if *X* is not diffident probably he is not malicious; *Y* perceives a positive social attitude in *X* and this elicits his goodwill; and so on.

(renounce to surveillance), and when through surveillance? A detailed model of how and why trust creates/increases trust is necessary in order to answer this question.

Should we make our autonomous agents (or our cyber-partners) more reliable and trustworthy 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, those subtle intertaction problems will certainly be relevant for computer-mediated human interaction and collaboration.

7.1.10 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 $Pr^*(achieve(g)) < Pr(achieve(g))$.
- Control might have the opposite effect than function (*kindA*), i.e. instead of improving performance, it might make performance worse. For example by producing anxiety in the trustee or by making him wast time and concentration for preparing or sending feedbacks (case in which $Trustworthiness^*(Y\ \tau) < Trustworthiness(Y\ \tau)$).
- It can produce a breakdown of willingness. Instead of reinforcing commitment and willingness, control can disturb it because of a bad reaction or rebellion, or because of delegation conflicts (Castelfranchi and Falcone, 1998) and need for autonomy; or because of the fact that distrust creates distrust (also in this case $Trustworthiness^*(Y\ \tau) < Trustworthiness\ (Y\ \tau)$).
- It can 'signal' (Section 6.3) to Y a lack of confidence and thus impact on (decrease) Y's self-confidence and self-esteem, negatively affecting his performance or commitment.

Here we mainly care about the bad effect of control on trust, which lets us see these dynamics. As trust virtuously creates trust, analogously the trust of Y in X, that can be very relevant for his motivation (for example in the case of exchange and collaboration), can decrease because X exibits not so much trust in Y (by controlling Y).

- X is too diffident, does this mean that X is malicious and Machiavellan? Since X suspects so much about the others would she herself be ready to deceive? Thus if X distrusts Y, Y can become diffident about X.
- Otherwise: X is too rigid, not the ideal person to work with.[15]
- Finally, if the agents rely on control, authority, norms they relax the moral, personal, or affective bonds, i.e. one of the strongest basis 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 of trust is unavoidable and vital.

[15] Control could also increase Y's trust in X, as a careful person, or a good master and boss, etc.

In sum, as for the dynamics of such a relation, we explained how:

- *X*'s Control over *Y* denounces and derives from a lack of *X*'s trust *in Y;*
- *X*'s Control over *Y* can increase *X*'s trust *in Y;*
- *X*'s Control over *Y* increases *X*'s trust in deciding to delegate to *Y* (her global trust);
- Control over *Y* by *X* can both increase and decrease *Y*'s trust in *X*; in case that 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 one);
- *X*'s control over *Y* improves *Y*'s performance, or makes it worse;
- *X*'s control over *Y* improves *Y*'s willingness, or makes him more demotivated.

7.1.11 Resuming the Relationships between Trust and Control

As we saw, relationships between trust and control are rather complicated. In this paragraph (see also Figure 7.9) we resume the different role that control can play with respect to trust.

In fact, as shown in Figure 7.9 the control can increase or decrease and in both the cases we can evaluate the potential influence on the two aspects of trust (strict and broad trust).

7.2 Adjusting Autonomy and Delegation on the Basis of Trust in *Y*

In this part we are going to analyze the complex scenario in which a cognitive agent (an agent with its own beliefs and goals) has the necessity to decide if and how to delegate/adopt a task to/for another agent in a given context. How much autonomy is necessary for a given task. How could this autonomy be changed (by both the trustor and the trustee) during the realization of the task. How *trust* and *control* play a relevant role in this decision and how important are their relationships and reciprocal influences.

Autonomy is very useful in cooperation (why someone should have an intelligent collaborator without exploiting its intelligence?) and even necessary in several cases (situatedness, different competence, local information and reactivity, decentralization, etc.), but it is also risky because of misunderstandings, disagreements and conflicts, mistakes, private utility, etc. A very good solution to this conflict is to maintain a high degree of interactivity *during* the collaboration, by providing *both* the man/user/client and the machine/delegee/contractor the possibility of taking the initiative in interaction and help (*mixed initiative* (Ferguson and Allen, 1998), (Hearst, 1999)) and of *adjusting* (Hexmoor, 2000) the kind/level of delegation and help, and the degree of autonomy run time.

We will analyze a specific view of autonomy which is strictly based on the notions of delegation and adoption (Castelfranchi and Falcone, 1998). In fact, in several situations the multi-agent plan, the cooperation between the delegating agent (*delegator*) and the delegated one (*delegee*), requires a strict collaboration and a control flow between the partners, in order to either maintain the delegator's trust or avoid breakdowns, failures, damages, and unsatisfactory solutions.

Software and autonomous agents will not only be useful for relieving human agents from boring and repetitive tasks; they will be mainly useful for situations where delegation and autonomy are necessary (*'strong dependence'*, Section 2.9) because the user/client/delegator does not have the local, decentralized and updated knowledge, or the expertise, or the

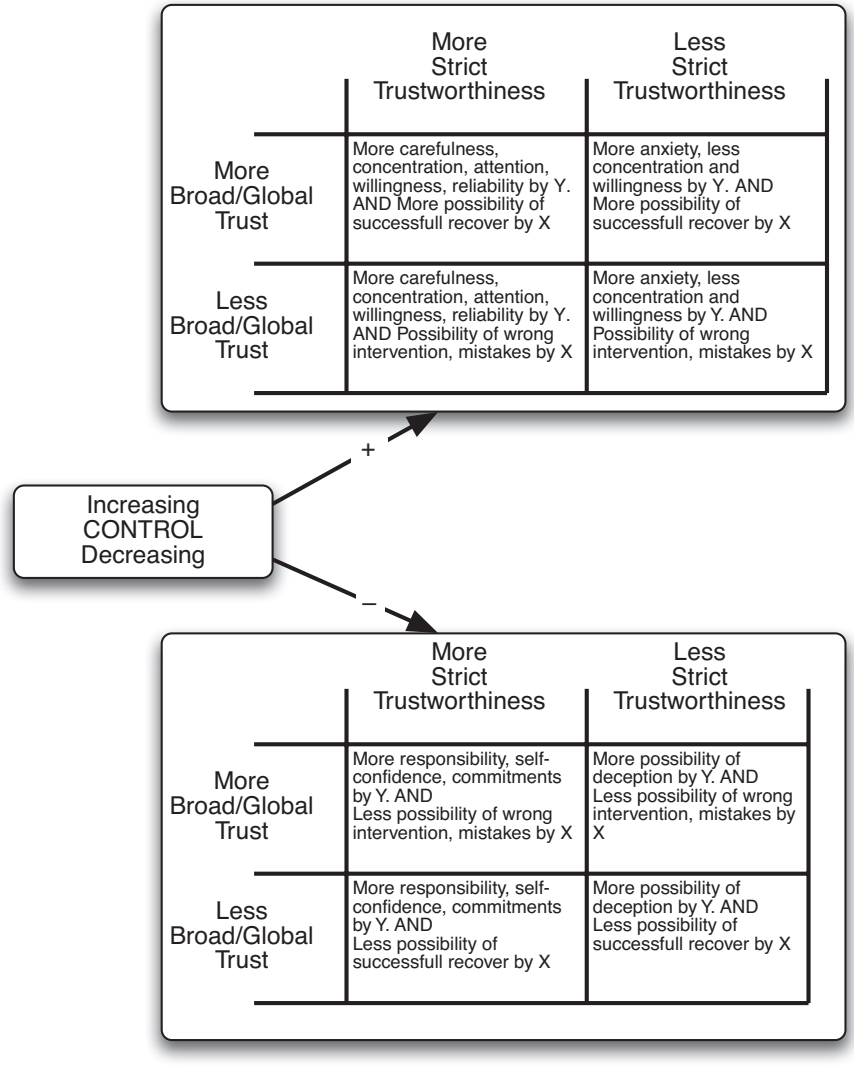

Figure 7.9 How Strict and Broad Trust change in relation with the Control's change

just-in-time reactivity, or some physical skill that requires some local control-loop. Thus autonomy and initiative are not simply optional features for *agents*, they are necessary requirements, and obligatory directions of study.

However, *control cannot be completely lost and delegation cannot be complete*, not only for reasons of confidence and trust, but for reasons of distribution of goals, of knowledge, of competence, and for an effective coordination. In this sense the possibility of controlling and adjusting the autonomy of the agents is becoming a growing and interesting field of research.

It has to be clear that this problem is central in any collaboration – between individuals, organizations, etc. – and that this theory is aimed at the general, not just for AI. Our claim in fact is that: *in designing how to adjust the level of autonomy and how to arrive at a dynamic level of control, it is necessary to have an explicit and general theory of (trust-based) delegation, which specifies different dimensions and levels of delegation, and relates the latter to the notion and the levels of autonomy*.

Thus, we propose our plan-based analysis (Pollack, 1990) of levels of delegation and levels of help, and discuss a related notion of autonomy. In several cases of collaboration among agents an *open delegation* is required, that is the delegation 'to bring it about that ...'. The agent is supposed to use its knowledge, intelligence, ability, to exert some degree of discretion.

Given that the knowledge of the delegating agent/user (client) concerning the domain and the helping agents is limited (both incomplete and incorrect) the 'delegated task' (the request or the elicited behavior) might not to be so useful for the delegator itself. Either the expected behavior is useful but cannot be executed, or it is useless or self-defeating, or dangerous for the delegator's other goals, or else there is a better way of satisfying the delegator's needs; and perhaps the helping agent is able to provide greater help with its knowledge and ability, going beyond the 'literally' delegated task. We will call *extension of help* or *critical help* this kind of help. To be really helpful this kind of agent must take the initiative of opposing (not for personal reasons/goals) the other's expectations or prescriptions, either proposing or directly executing a different action/plan. To do this it must be able to recognize and reason with the goals, plans and interests of the delegator, and to have/generate different solutions.

Open delegation and over/critical help distinguish a *collaborator* from a simple tool, and presupposes intelligence and autonomy (discretion) in the agent. However, of course, there is a trade-off between pros and cons both in *open delegation* and in *extended(critical)-help*: the more intelligent and autonomous the delegee (able to solve problems, to choose between alternatives, to think rationally and to plan) the less *passively obedient* it is.[16] So, possible conflicts arise between a client and its contractor; conflicts which are due either to the intelligence and the initiative of the contractor or to an inappropriate delegation by the client. We are interested here only in the conflicts originating from the agent's willingness to collaborate and to help the other in a better and more efficient way: a kind of *collaborative conflict*. We do not consider the contractor's *selfish reasons* for modifying delegation (because the nature of the conflict, negotiation, etc. would be different).[17]

It is worth specifying that this work is aimed at providing a theoretical framework, i.e. the conceptual instruments necessary for analyzing and understanding interaction with autonomous entities. As has just been said, we assume that this framework is useful not only for organization theory or management, but also for a principled engineering, i.e. for getting systems designed not only on the basis of empirical data and practical experience, but also on the basis of a more complete view and typology, and of some prediction and explanation. The role of the dynamics of trust (Chapter 6) and control in this conflict and in the adjustment of the level of *Y*'s autonomy is clear.

We also suggest some criteria about when and why to adjust the autonomy of an agent, and preliminary hints about necessary protocols for adjusting the interaction with agents.

[16] Obviously a very autonomous but stupid agent is even worse.

[17] In this chapter there is at least one of these reasons that should be taken into account: when the contractor/trustee adjusts the delegation for having more autonomy.

7.2.1 The Notion of Autonomy in Collaboration

For the purpose of this book we use a practical and not very general notion of autonomy.[18] In particular, we refer to the *social* autonomy in a *collaborative* relationship among agents. We distinguish between:

- a *meta-level autonomy* that denotes how much the agent is able and in condition to negotiate over the delegation or indeed to change it (in this regard, a slave, for example, is not autonomous: he cannot negotiate or refuse);
- a *realization autonomy*, that means that the agent has some discretion in finding a solution to an assigned problem, or a plan for an assigned goal.

Both are forms of goal-autonomy, the former at the higher level, the latter at the sub-goals (instrumental) level. For definition of different kinds of autonomy, including some of the dimensions we consider, see also (Huhns and Singh, 1997).

The lower the control of the client/trustor (monitoring or intervention) on the execution, the more autonomous is the contractor. In this context then, *autonomy means the possibility of displaying or providing an unexpected behavior (including refusal) that departs from the requested (agreed upon or not) behavior. The autonomous agent can be either entitled or not to perform such an unexpected behavior*.[19]

7.2.2 Delegation/Adoption Theory

We introduced the delegation notion in Section 2.9. Here we use that concept, integrating it with the notion of adoption and developing the theory of adjustable autonomy.

[18] We do not consider here some important aspects of autonomy (that could be adjusted) like the agent's independence or self-sufficiency. For an analytical discussion on the notion of *autonomy* in agents and for a more principled definition, see (Martin and Barber, 1996), (Castelfranchi, 2000b), and (Castelfranchi, 1995).

[19] In this book we do not discuss in detail another very important distinction between:

- being *practically in condition of* doing something (refusing, negotiating, changing and doing something else), i.e. what we would like to call ≪practical possibility≫; and
- being *deontically in condition of* doing something, i.e. to be entitled, permitted in the strong sense, i.e. the ≪deontic possibility≫.

An agent can have the former without the latter, or vice versa (see (Castelfranchi, 2000b)). In fact, there are two kinds of lack of power (hence, of dependence and autonomy): one based on practical conditions, the other based on deontic conditions. In deontic autonomy, an agent is permitted to do/decide/ interpret/ infer/ etc. Not only is it practically able and in condition to, but it can do this without violating a social or legal norm, or the user/designer prescriptions. As there are two kinds of autonomy there are two kinds of 'empowerment' (giving autonomy): deontic empowerment versus practical, material empowerment (Jones and Sergot, 1996). Therefore, an additional dimension of *adjustment* should be taken into account that is, the deontic one. The delegator (or the delegee) can attempt to modify (restrict or enlarge) either what the delegee is practically and actually able to do independently of the other, or what it is entitled to do. For example when a delegee re-starts negotiation, instead of directly modifying the task, it is implicitly asking some sort of permission, or agreement. Obviously enough, in strong delegation (contract relation, see later) the assignment of a task τ to the delegee implicitly entails giving it the permission to do τ. Adjusting the entitled space of freedom, or adjusting the practical space of freedom, is an interesting difference, but we cannot examine it in this book. Notice that this theory would imply the same plan-based dimensions of delegation and help.

Formal Constructs

Several formal constructs are needed in the following. Let $Act = \{\alpha_1, \ldots, \alpha_n\}$ be a set of *actions*, and $Agt = \{Ag_1, \ldots, Ag_m\}$ a set of *agents*. The *general plan library* is $\Pi = \Pi^a \cup \Pi^d$, where Π^a is the abstraction hierarchy rule set and Π^d is the decomposition hierarchy rule set. An action $\alpha' \in Act$ is called *elementary action* in Π if there is no rule r in Π such that α' is the left part of r. We will call *BAct* (*Basic Actions*) the set of elementary actions in Π and *CAct* (*Complex Actions*) the remaining actions in *Act*.

Given α_1, α_2 and Π^d, we introduce the *Dom-c*$(\alpha_1\ \alpha_2)$ operator to say that α_1 *dominates* α_2 (or α_2 *is dominated by* α_1) in Π^d: *Dom-c*$(\alpha_1\ \alpha_2)$=*True* if there is a set of rules $(r_1, \ldots, r_m)$ in Π^d, such that: $(\alpha_1 = Lr_1) \wedge (\alpha_2 \in Rr_m) \wedge (Lr_i \in Rr_{i-1})$, where: Lr_j and Rr_j are, respectively, the left part and the right part of the rule r_j and $2 \leq i \leq m$ (in the same way it is possible to define the *Dom-a*$(\alpha_1\ \alpha_2)$ operator considering the abstraction hierarchy rule set Π^a). We denote Π_{Agx} as the Ag_x's plan library, and Act_{Agx}, the set of actions known by Ag_x. The set of irreducible actions (through decomposition or specification rules) included in Π_{Agx} is composed of two subsets: the set of actions that Ag_x believes to be elementary actions ($BAct_{Agx}$) and the set of actions that Ag_x believes to be complex but for which it has no reduction rules ($NRAct_{Agx}$: *Non Reduced actions*). Then $BAct_{Agx}$ is included in *Act* and possibly $BAct_{Agx}$ is included or coincides with *BAct*. In fact, given an elementary action, an agent may (or may not) know the body of that action. We define S_{Agx} as the *skill set* of Ag_x, the actions in $BAct_{Agx}$ whose body is known by Ag_x (action repertoire of Ag_x).[20] We call R the operator that, when applied to an action α, returns the set of the *results* produced by α.

Definition of Delegation and Adoption

The domain of MAS, collaboration (Haddadi, 1996), and teamwork are already familiar with the notion of delegation. However, our analysis is grounded on more basic notions (Hexmoor, 2000). In addition, our delegation theory is not limited to explaining and modeling interpersonal relationships; the basic concepts of our definition also apply to (and are necessary even if not sufficient for) other important concepts such as:

- *institutional delegation*, in which the delegator transfers to the delegee not just some task but also some right, obligation, responsibility, power and so on (Jones and Sergot, 1996). Of course, this notion is richer than our basic concept (see below).
- *roles and prescriptions in organizations*, roles can be analyzed also as sets of delegated tasks (Castelfranchi and Falcone, 1997).

In our model, *delegation and goal adoption are characterized in terms of the particular set of mental states (cognitive ingredients) of the agents involved in the interaction*. Informally, *in*

[20] In sum, an agent Ag_x has its own plan library, Π_{Agx}, in which some actions ($CAct_{Agx}$ and $NRAct_{Agx}$) are complex actions (and it knows the reduction rules of $CAct_{Agx}$) while some other actions ($BAct_{Agx}$) are elementary actions (and it knows the body of a subset - S_{Agx} - of them).

delegation (reliance) an agent Ag_1 needs or likes an action of another agent Ag_2 and includes it in its own plan (see Section 2.9).

In adoption (help) an agent Ag_2 acquires and has a goal as (long as) it is the goal of another agent Ag_1, that is, Ag_2 has the goal of performing an action because this action is included in the plan of Ag_1. So, also in this case Ag_2 plays a part in Ag_1's plan (sometimes Ag_1 has no plan at all but just a need, or a goal) since Ag_2 is doing something for Ag_1.

We consider the action/goal pair $\tau=(\alpha,g)$ as the real object of delegation,[21] and we called it a 'task'. Then by τ, we will refer to the action (α), to its resulting world state (g), or to both. We introduce an operator of delegation with three parameters:

$$Delegates(Ag_1\ Ag_2\ \tau) \tag{7.7}$$

where Ag_1, Ag_2 are agents and $\tau=(\alpha,g)$ is the task. This means that Ag_1 delegates the task τ to Ag_2. In analogy with delegation we introduce the corresponding operator for adoption:

$$Adopts(Ag_2\ Ag_1\ \tau) \tag{7.8}$$

This means that Ag_2 adopts the task τ for Ag_1: Ag_2 helps Ag_1 by caring about τ.

Dimensions of Delegation and Adoption

We consider three main dimensions of delegation/adoption: *interaction-based*, *specification-based*, and *control-based* types of delegation/adoption (Castelfranchi and Falcone, 1998). Let us analyze these cases more in detail.

- *Interaction-based types of delegation.* Three general cases may be given: *weak*, *mild* and *strong delegation*. They represent different degrees of strength of the established delegation. In the following we synthesize (more formal details can be find in Section 2.9) the mental ingredients of trust in the different delegation actions.

W-Delegates is the operator for representing *weak delegation.* So the expression:

W-Delegates(Ag_1 Ag_2 τ) represents the *necessary* mental ingredients for Ag_1 trusting Ag_2 on the task τ, shown in Figure 2.11 and resumed in a less formal way in Table 7.1.

We consider in Table 7.1 (a, b, c, and d) what the agent Ag_1 views as a '*Potential for relying on*' agent Ag_2, its *trust* in Ag_2; and (e and f) what Ag_1 views as the '*Decision to rely on*' Ag_2.

[21] We assume that *delegating an action necessarily implies delegating some result of that action* (i.e. expecting some results from Ag_2's action and relying on it for obtaining those results). Conversely, *to delegate a goal state always implies the delegation of at least one action (possibly unknown to Ag_1) that produces such a goal state as a result* (even when Ag_1 asks Ag_2 to solve a problem, to bring it about that g without knowing or specifying the action, Ag_1 necessarily presupposes that Ag_2 should and will do some action and relies on this).

Table 7.1 Mental Ingredients in Weak-Delegation (pseudo-formal description)

a) The achievement of τ is a *goal* of Ag_1.
b) Ag_1 believes that there exists another agent Ag_2 that has the *power of* achieving τ.
c) Ag_1 believes that Ag_2 will achieve τ in time and by itself (without Ag_1's intervention).
(if Ag_2 is a cognitive agent, Ag_1 believes that Ag_2 *intends* to achieve τ.).
d) Ag_1 *prefers*[22] to achieve τ through Ag_2.
e) The achievement of τ through Ag_2 is the choice (goal) of Ag_1.
f) Ag_1 has the goal (*relativized* (Cohen and Levesque, 1987) to (e)) of not achieving τ by itself.

Table 7.2 Mental Ingredients in Mild-Delegation (pseudo-formal description)

a′≡a; b′≡b; d′≡d; e′≡e; f′≡f; (referring to *a*, *b*, *d*, *e*, and *f* as described in Table 7.1)
c′) Ag_1 does not believe that Ag_2 will achieve τ by itself (without Ag_1's intervention).

g′) Ag_1 believes that if Ag_1 realizes an action α' then it is be more probable that Ag_2 intends τ.
But Ag_2 does not adopt $Ag_1's$ goal that Ag_2 intends τ.

h′) Ag_1 intends to do α' relativized to (e′).

We consider 'Potential for relying on' and 'Decision to rely on' as two constructs temporally and logically related to each other .[23]

M-Delegates is the operator for representing *mild delegation*.

M-Delegates(Ag_1 Ag_2 τ) represents the necessary mental ingredients of trust shown in Figure 2.12 and resumed in less formal way in Table 7.2.

We consider in Table 7.2 (a', b', c', d' and e') what agent Ag_1 views as a *'Potential for relying on'* agent Ag_2; and (f', g' and h') what Ag_1 views as the *'Decision to rely on'* Ag_2.[24]

S-Delegates is the operator for representing *strong delegation*. So the expression *S-Delegates(Ag_1 Ag_2 τ)* represents the *necessary* mental ingredients of trust as shown in Figure 2.13 and resumed in less formal way in Table 7.3:

We consider in Table 7.3 (a'', b', c'', d'' and e'') what agent Ag_1 views as a *'Potential for relying on'* agent Ag_2; and (f'', g'' and h'') what Ag_1 views as the *'Decision to rely on'* Ag_2.

For a corresponding analysis of adoption, and for how the kind of interaction between client and contractor influences the adoption itself see (Castelfranchi and Falcone, 1998).

[22] This means that Ag_1 believes that either the achievement of τ or a broader goal *g'* that includes the achievement of τ, implies Ag_1 to be dependent on Ag_2. Moreover (d) implies Ag_1's goal that Ag_2 achieves τ.

[23] As for weak delegation it is interesting to analyze the possibilities of Ag_2's mind. We should distinguish between two main cases: Ag_2 knows $W - Delegates(Ag_1 Ag_2\ \tau)$ and Ag_2 does not know $W - Delegates(Ag_1 Ag_2 \tau;)$. In other words, a weak delegation is possible even if the delegee knows it. Either this knowledge has no effect (the achievement of Ag_1's goal is just a side-effect known by Ag_2) or this knowledge changes Ag_2's goal: Ag_2 can either arrive at spontaneous and unilateral help or to a reactive, hostile attitude.

[24] In analogy with what we have said in weak delegation, also in mild delegation we should distinguish between two main cases about the possible mental states of Ag_2: Ag_2 knows *M-Delegates(Ag_1 Ag_2 τ)* and Ag_2 does not know *M-Delegates(Ag_1 Ag_2 τ)*. So, it is possible to have a mild delegation even if the delegee knows it and if consequently it changes its own behavior to favor or to hamper the success of it.

Table 7.3 Mental Ingredients in Strong-Delegation (pseudo-formal description)

a″≡a; b″≡b; c″≡c d″≡d; e″≡e; f″≡f; (referring to *a, b, c, d, e*, and *f* as described in Table 7.1)
g″) Ag_1 believes that if Ag_1 realizes an action α' there will be this result: Ag_2 will intend τ as the consequence of the fact that Ag_2 adopts $Ag_1's$ goal that Ag_2 intends τ (in other words, Ag_2 will be socially committed to Ag_1).
h″) Ag_1 has the goal (*relativized* to (e')) of not achieving τ by itself.

- *Specification-based types of delegation/adoption*. How is the task specified in delegation and how does this specification influence the contractor's autonomy? The object of delegation/adoption (τ) can be minimally specified (*open delegation*), completely specified (*closed delegation*) or specified at any intermediate level. Let us consider two cases:
  i) *Merely Executive (Closed) Delegation:* here the client (or the contractor) believes it is delegating (adopting) a completely specified task; what Ag_1 expects from Ag_2 is just the execution of a sequence of elementary actions (or what Ag_2 believes Ag_1 delegated to it is simply the execution of a sequence of elementary actions).[25]
  ii) *Open Delegation:* when the client (contractor) believes it is delegating (adopting) a non completely specified task: either Ag_1 (Ag_2) is delegating (adopting) an abstract action, or it is delegating (adopting) just a result (i.e. a state of the world).[26] Ag_2 can realize the delegated (adopted) task by exerting its autonomy. We can have several possible levels of openness of the delegated (adopted) task.
- *Control-based types of Delegation*. In this case we distinguish the delegation on the basis of the level of control it implies. At one extreme we have 'full control' (in fact the delegee is always under control during the realization of the delegated task) while at the other extreme we have 'no control' (in fact the delegee is never under control during the realization of the delegated task). As we have seen there are two main kinds in the control dimension: monitoring and intervention. Both have to be considered as influencing the delegation (we do not consider here a more detailed analysis of their different influences on delegation).

In Figure 7.10 we summarize the three main dimensions of delegation: each characterizes the variability of delegation action. The delegee's autonomy decreases towards the origin of the Cartesian space within the solid. Each of these dimensions implies, in fact, a specific aspect of the delegee's autonomy about the task.

7.2.3 The Adjustment of Delegation/Adoption

Run-Time Adjustment

We can consider the adjustment of autonomy (the revision of delegation) in three different time periods:

[25] More formally, either $\alpha \in S_{Ag1}$, or $\alpha \in BAct_{Ag1}$ ($\alpha \in S_{Ag2}$, *or* $\alpha \in BAct_{Ag2}$,), or g is the relevant result of α and $\alpha \in S_{Ag1}$ or $\alpha \in BAct_{Ag1}$ ($\alpha \in S_{Ag2}$, *or* $\alpha \in BAct_{Ag2}$,).

[26] More formally, either $\alpha \in CAct_{Ag1}$, or $\alpha \in NRAct_{Ag1}$ (either $\alpha \in CAct_{Ag2}$, or $\alpha \in NRAct_{Ag2}$); and also when g is the relevant result of α and $\alpha \in CAct_{Ag1}$ or $\alpha \in NRAct_{Ag1}$ ($\alpha \in CAct_{Ag2}$, *or* $\alpha \in NRAct_{Ag2}$).

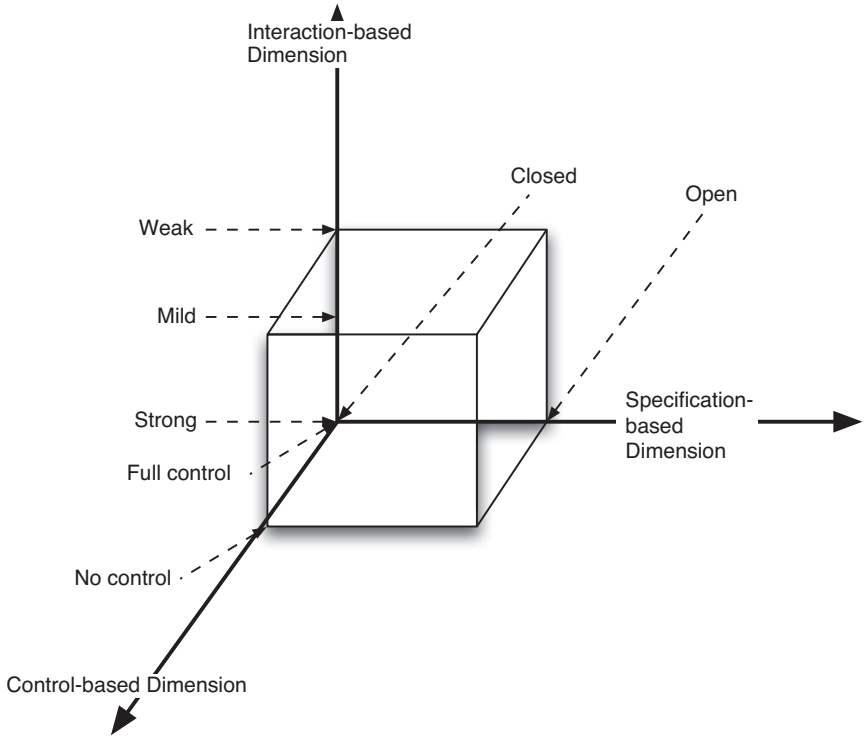

Figure 7.10 The three main dimensions of delegation. (Reproduced by Permission of © 2001 IEEE)

i) *After the delegation event,* but *before the execution of the task;*
ii) *Run-time,* with *work-in-progress;*
iii) *At the end of the performance* and the realization of the task; in this case the adjustment of autonomy will have an effect only on a future delegation (we can consider this case as a case of *learning*).

We will focus here on *run-time* adjustment, this is particularly important in human-machine interaction and in multi-agent cooperation, and call it simply *adjustment*. We will first examine the problem of adjustable autonomy *in a broad sense*, i.e. as adjusting the level and the kind of delegation/adoption (in our terminology *delegation conflicts* (Castelfranchi and Falcone, 2000c)). We claim that this is the right theoretical frame for understanding also the adjustment of autonomy in a strict sense, since *any autonomy adjustment requires a delegation adjustment*, but not vice versa (see next section).

In the following, we will analyze the general reasons for delegation/adoption adjustment. Let us here consider the taxonomy of the adjustments (some of which will be neglected because meaningless), their nature and their importance. Each of the possible adjustments is *bilateral*, i.e. either the client or the contractor can try to modify the previous delegation.

Table 7.4 Adjustments with respect to the interaction dimension

Line Number	*Agent that has the initiative of the adjustment*	*Starting State*	*Final State*
1	*Delegator*	*Weak delegation*	*Mild delegation*
2	*Delegator*	*Weak delegation*	*Strong delegation*
3	*Delegator*	*Mild delegation*	*Strong delegation*
4	*Delegator*	*Mild delegation*	*Weak delegation*
5	*Delegator*	*Strong delegation*	*Weak delegation*
6	*Delegator*	*Strong delegation*	*Mild delegation*
7	*Delegee*	*Weak delegation*	*Mild delegation*
8	*Delegee*	*Weak delegation*	*Strong delegation*
9	*Delegee*	*Mild delegation*	*Strong delegation*
10	*Delegee*	*Strong delegation*	*Mild delegation*
11	*Delegee*	*Strong delegation*	*Weak delegation*
12	*Delegee*	*Mild delegation*	*Weak delegation*
13	*Delegator*	*Weak adoption*	*Strong adoption*
14	*Delegator*	*Strong adoption*	*Weak adoption*
15	*Delegee*	*Weak adoption*	*Strong adoption*
16	*Delegee*	*Strong adoption*	*Weak adoption*

Source: Reproduced by Permission of © 2001 IEEE.

Delegation/Adoption Adjustments with Respect to the Interaction Dimension

As described in Table 7.4 there are (with respect to the interaction dimension) several possibilities of adjustment; they are determined by:

- the *agent* who has the initiative of the adjustment;
- the *starting state* (the kind of delegation or adoption acting in that given instant and that the *agent* intends to modify);
- the *final state* (the kind of delegation or adoption to which *agent* intends to arrive).

A few cases shown in Table 7.4 deserve some comments.

- *Line 1*: can be inferred from the difference between the mental ingredients of weak (see Table 7.1) and mild (see Table 7.2) delegation: in fact, c is replaced by c', g' and h'. In other words, Ag_1 does not believe that Ag_2 will achieve τ without any influence and so decides to realize an action α' that could produce this influence. In this case there is still no social commitment (Castelfranchi, 1996) by Ag_2: Ag_2 does not adopt Ag_1's goal that Ag_2 intends τ . In this case there is no sufficient trust and the trustor decides, for achieving the task, to introduce additional influences on the trustee.
- *Line 2*: g'' and h'' are added beliefs. In other words, Ag_1 tries to achieve τ through a social commitment of Ag_2: for this it realizes α'.
- *Lines 8 and 9*: could represent the willingness of Ag_2 to convert Ag_1's exploitation into a clear social commitment between them.
- *Lines 13-14*: are linked with the initiative of Ag_1 in the case in which Ag_1 is aware of Ag_2's adoption.

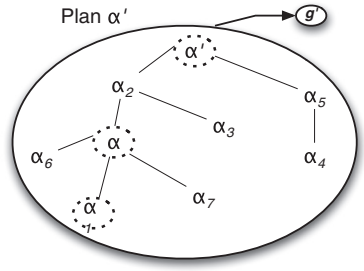

Figure 7.11 Composition Relationship

Delegation/Adoption Adjustments with Respect to the Specification Dimension

Also in these cases we must consider the intervention of both the client/trustor and the contractor/trustee with regard to delegation and adoption, respectively. Before analyzing the different cases included in this dimension, we show (also graphically) the meaning of the relationship about action composition (we called it *Dom-c*: *Dom-c*(α' α) defines this relationship between α' and α, where α is a component action of α').

Given Figure 7.11 we can say that the plan (complex action) α' gives the results for achieving the goal g'. This complex action is constituted from a set of different actions related with each other by the composition relationship (for example: *Make-Dinner* is constituted at a first level from *Buy-food*, *Prepare-food* and *Cook-food*; each of this action is, in its turn, constituted from other elementary or complex actions, and so on).

In fact, each of the actions shown in Figure 7.11 produces (temporary or final) results: see Figure 7.12. Temporary results will not be present in the results produced by the plan at the end of its execution (while final results will be present).

Delegee's Adjustments

The reasons for the delegee's adjustments can be of different nature and related to selfish or cooperative goals. In any case, these reasons cannot be irrespective of the evaluation of his own

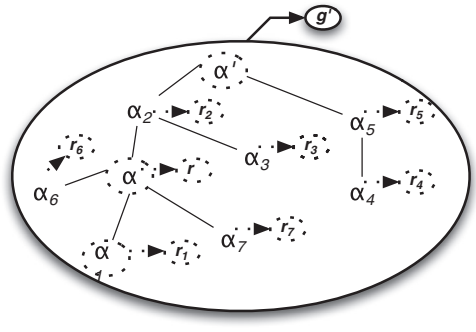

Figure 7.12 Composition Relationship with the evidence of the component action results

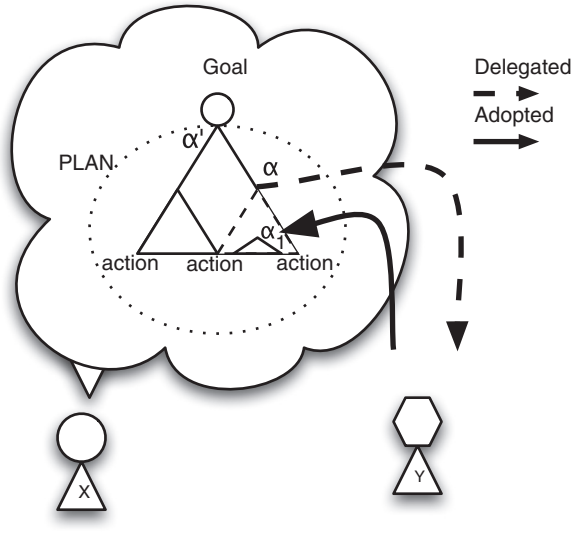

Figure 7.13 Sub-help

attitudes in that specific task (a sort of selftrust). Suppose that *Delegates*(Ag_1 Ag_2 τ) and τ is included in a more general Ag_1's plan aimed at achieving goal g' through a complex action α'. Moreover, *Dom-c*(α' α), $\tau=(\alpha,g)$, and $\tau'=(\alpha',g')$. We have three main delegee's adjustments:

- Reduction of help

  Here the *delegee provides less help on* τ *than delegated*. If

$$Adopts(Ag_2\ Ag_1\ \tau_1) \wedge Dom\text{-}c(\alpha\ \alpha_1) \tag{7.9}$$

  with $\tau_1=(\alpha_1,g_1)$, the delegee reduces the task to a subpart of the requested one (see Figure 7.13). For example Ag_1 delegates Ag_2 to prepare a main course for a dinner and bring it to her house. Ag_2 only buys the ingredients but does not cook them.

  A sub-help is not necessary a help (although lower than expected). Maybe the realized action and the achieved subgoal are completely useless (in a plan the coordination among the actions and their results are also very important).

  For example, in the case of the previous example, maybe the ingredients cannot be cooked at Ag_1's house because of a problem with the kitchen.

- Extension of help

  Here the delegee provides more help on τ than delegated. If

$$Adopts(Ag_2\ Ag_1\ \tau_1) \wedge Dom\text{-}c(\alpha_1\ \alpha) \wedge (Dom\text{-}c(\alpha'\ \alpha_1)\ OR\ (\alpha' \equiv \alpha_1)) \tag{7.10}$$

  with $\tau_1=(\alpha_1,g_1)$; the delegee goes beyond what has been delegated by the client without changing the delegator's plan (Figure 7.14). In fact, the delegee chooses a task that satisfies a

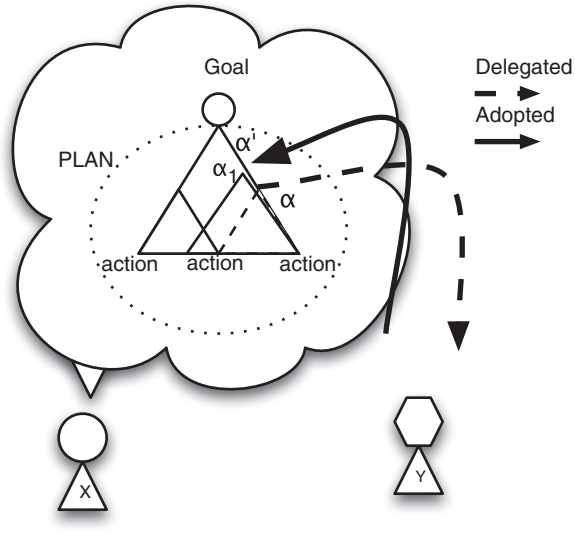

Figure 7.14 Over-help

higher level task (within the general delegator's intended plan) compared with the delegated task.

An example for this case is when Ag_1 delegates Ag_2 to cook 'spaghetti with pesto' but Ag_2 prepares the whole dinner.

- Critical help

It is the case in which the *delegee provides a qualitatively different action/help than expected* (what has been delegated). Let us analyze some subcases:

- *Simple critical help*

$$Adopts\,(Ag_2\ Ag_1\ \tau_x) \wedge g \in R(\alpha_x) \tag{7.11}$$

with τ_x=(α_x,g); the delegee achieves the goal(s) of the delegated plan/action, but it changes that plan/action (Figure 7.15). An example of *Simple Critical help* is when Ag_1 delegates Ag_2 to cook 'spaghetti with pesto' but Ag_2 has already bought cooked 'spaghetti with pesto' (he changes his own actions but the final result of them is supposed to be the same).

- *Critical overhelp*

$$Adopts\,(Ag_2\ Ag_1\ \tau_x) \wedge g_1 \in R(\alpha_x) \wedge Dom\text{-}c(\alpha_1\ \alpha) \wedge (Dom\text{-}c(\alpha'\ \alpha_1)\ OR\ (\alpha' \equiv \alpha_1)) \tag{7.12}$$

with $\tau_1 = (\alpha_1, g_1)$ $\tau_x = (\alpha_x, g_1)$, $R(\alpha_x)$ the set of results produced by α_x (see Figure 7.16); the delegee implements both a simple critical help and an extension of help (it chooses a task that

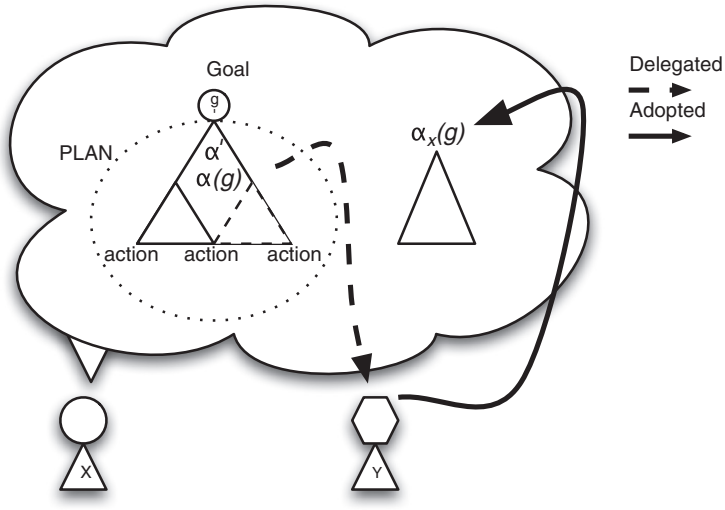

Figure 7.15 Simple Critical Help

satisfies a higher level task with respect to the task delegated, and achieves the goal(s) of this higher task, while changing the expected plan).

An example of *Critical Overhelp* is when Ag_1 delegates Ag_2 to cook 'spaghetti with pesto' (a subplan of preparing a dinner, thinking of cooking the other courses herself and in this way achieving the goal of offering a dinner to some old friends). But Ag_2 reserves a famous

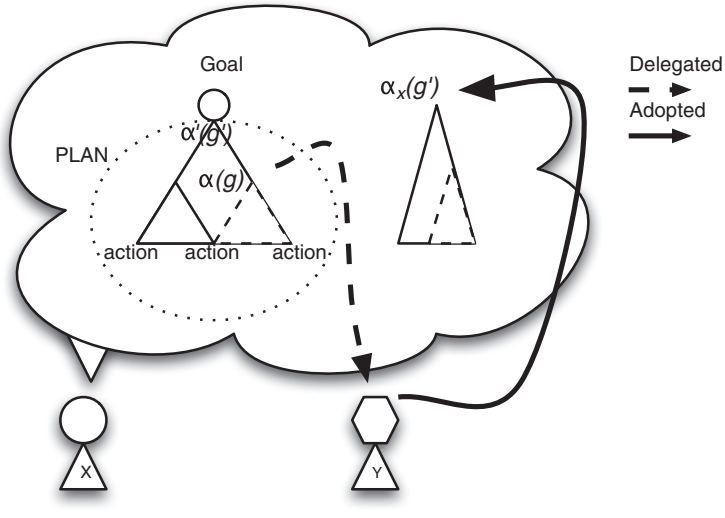

Figure 7.16 Critical Over-Help

restaurant in the town (in this way conserving Ag_1's goal of offering a dinner to some old friends).

- *Hyper-critical help*

$$Adopts\,(Ag_2\ Ag_1\ \tau_1) \wedge (g_1 \neq g') \wedge (g_1 \neq g) \wedge (g_1 \in I_{Ag1}) \tag{7.13}$$

with $\tau_1 = (\alpha_1, g_1)$, and I_{Ag1} is the set of interests of Ag_1. Ag_2 adopts goals or interests of Ag_1 that Ag_1 themselves did not take into account: by doing so, Ag_2 neither performs the delegated action/plan nor achieves the results that were delegated.

A typical example of *Hyper-critical Help* is when Ag_1 asks Ag_2 for a cigarette and Ag_2 says to Ag_1 'you must not smoke'. In this way Ag_2 is dictating an interest of Ag_1 (to be healthy).

Delegator's Adjustment

The reasons for the delegator's adjustments can be different and related to selfish or cooperative goals. In any case, these reasons cannot be irrespective of the delegator's trust in the trustee about that specific task. Suppose that $Delegates(Ag_1\ Ag_2\ \tau)$, and that Ag_1 intends to change that delegation. Suppose also that Ag_1 is achieving goal g' through plan τ', with $Dom\text{-}c(\alpha'\ \alpha)$. We can have five main delegator's adjustments:

- Reduction of delegation

It is the case in which there is a new delegation:

$$Delegates(Ag_1\ Ag_2\ \tau_1) \wedge Dom\text{-}c(\alpha\ \alpha_1) \tag{7.14}$$

with $\tau_1 = (\alpha_1, g_1)$, the delegator adjusts the original delegation, by *reducing the task that the contractor must realize* (the client reduces the task to a subpart of the previous requested task).

For example, the delegator no longer trusts the trustee to complete the more complex action (see Figure 7.17).

- Extension of delegation

$$Delegates\,(Ag_1\ Ag_2\ \tau_1) \wedge Dom\text{-}c(\alpha_1\ \alpha) \wedge (Dom\text{-}c(\alpha'\ \alpha_1)\ OR\ (\alpha' \equiv \alpha_1)) \tag{7.15}$$

with $\tau_1 = (\alpha_1, g_1)$, the delegator adjusts its delegation in such a way that its *new request goes beyond what has been originally delegated without changing the previous plan* (see Figure 7.18).

- Modification of delegation

In an analogy with the delegee's adjustments, which consists of four subcases (modification of the previous delegated task just changing the previous goal; modification of the previous

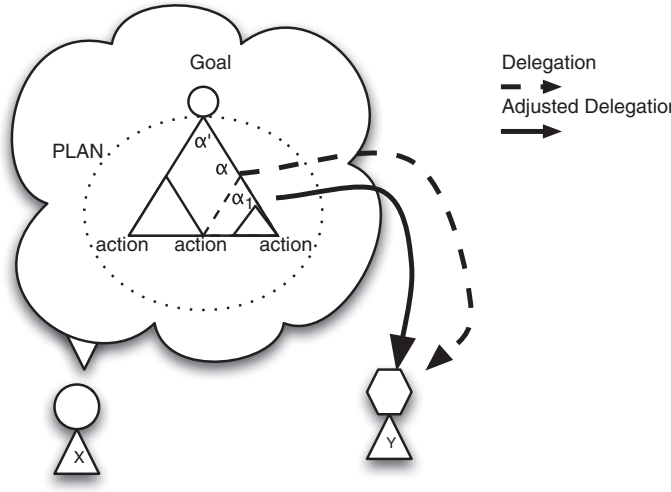

Figure 7.17 Riduction of Delegation

delegated task considering an over-goal and changing the plan to obtain that over-goal; modification of the previous delegated task considering a sub-goal and changing the plan to obtain that sub-goal; modification of the previous delegated task changing both the plan and the goal).

- Openness of delegation

$$Delegates\,(Ag_1\ Ag_2\ \tau_x) \wedge Dom\text{-}a(\alpha_x\ \alpha) \tag{7.16}$$

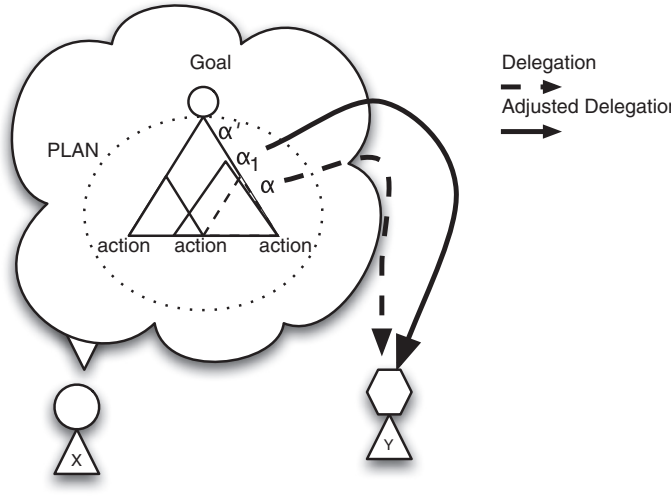

Figure 7.18 Extension of Delegation

In other words, the delegator adjusts their own delegation so that the new delegated plan is more abstract.

For example Ag_1 changes her delegation to Ag_2 from 'prepare a dinner composed of spaghetti with pesto' and 'chicken with french fries' to 'I need something to eat' (Ag_1 can cook a different dinner or can buy something to eat or can invite Ag_1 to a restaurant, and so on).

- Closing of delegation

$$Delegates\,(Ag_1\ Ag_2\ \tau_x) \wedge Dom\text{-}a(\alpha\ \alpha_x) \tag{7.17}$$

In other words, the delegator adjusts its own delegation so that the new delegated plan is more specified.

7.2.4 Channels for the Bilateral Adjustments

For adjusting delegation and help, channels and protocols are necessary. As we have said, on the trustor/client's side, they are useful for monitoring (reporting, observing, inspecting), and intervention (instructions, guidance, helps, repair, brake); on the trustee/delegee's side it is useful to have some space for discretion and practical innovation. For both client and contractor, are useful channels and protocols for communication and re-negotiation during the role-playing and the task execution.

The Trustor's Side

As we have already written, Ag_1 must have some form of *control* on and during Ag_2's task realization, otherwise no adjustment is possible. Obviously, the feedback, i.e. monitoring, is useful for a run-time adjustment and must be provided in time for the intervention. In general, the feedback activity is the precondition for an intervention.

In multi-agent systems, in order to guarantee agents a dynamic adjustment of delegation and their mixed initiative we have to provide such an *infrastructure*, while in human-computer interaction we have to provide the user with those two channels.

When Ag_2 has the initiative (it is starting with an adoption action), if Ag_1 wants to change this adoption it needs a *communication channel* with Ag_2.

The Trustee's Side

Ag_2 can run-time adjust the delegation of Ag_1 (and its own autonomy) if it is in condition of either:

a) having a *communication channel* for *(re-)starting negotiation* by offering/proposing a different level of help to Ag_1; or
b) having enough practical freedom to *directly change* the action (this condition should be by definition a characteristic of autonomous agents).

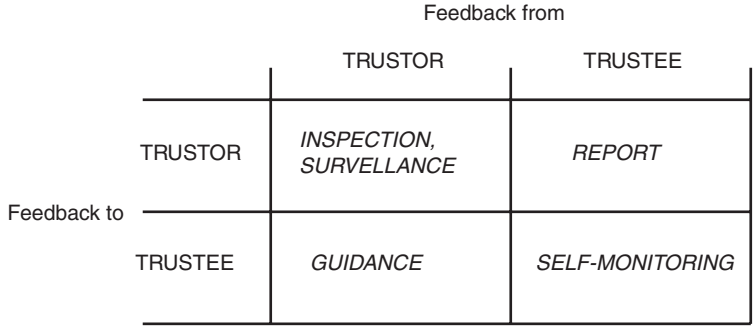

Figure 7.19 How feedback determines different kinds of control. (Reproduced by Permission of © 2001 IEEE)

Trustees should not necessarily negotiate or give advice in order to change their delegated tasks; they might have full initiative. This entails meta-level autonomy. Of course, the trustee must also have feedback about its own execution, but this is true in general for goal directed actions.

To sum up, if an agent has the initiative of a delegation/adoption then, in order to adjust that initiative, it is not obliged to communicate with the other agent. As for the necessary *feedback for adjustment* we can distinguish between: *inspection*, *report*, *guidance*, and *self-monitoring* (Figure 7.19).

Considered the kinds of intervention action showed in Section 7.1.2, we can say that each of these interventions could be realized through either a *communication act* or a *direct action* on the task by the intervening agent (Table 7.5).

7.2.5 Protocols for Control Adjustments

Starting from our model it is also possible to identify some guidelines for designing interaction in human-machine interaction or in multi-agent systems. For example, our model makes it

Table 7.5 Different kinds of client intervention

	Client's message	Client's direct action
Stopping the task	Stop	Stopping intervention
Substitution	I do it	It realizes an action of the task
Correction of delegation	Change that action with this other	It introduces constraints such that an action is changed with another
Specification of delegation	Make that plan in this way	It introduces constraints such that a plan is specified
Repairing of delegation	Add this action of the task	It introduces constraints such that a new action must be realized to favor success of the task

Source: Reproduced by Permission of © 2001 IEEE.

clear that agent control requires either communication or a simple action-perception loop. On the one hand, Ag_1 can monitor Ag_2 without any communication (without sending any specialized message or signal), by simply observing it; at the same time Ag_1 can also influence Ag_2 by physical/practical actions and interventions on Ag_2 or on the world. For example Ag_1 can brake and stop Ag_2. On the other hand, Ag_1 can monitor Ag_2 thanks to messages sent by Ag_2 to Ag_1 (reports), and can influence Ag_2 by sending them messages (instructions, warnings, etc.).

Examples of Monitoring

Let us show you how the *monitoring* actions can be expressed in praxis and in communication.

PRAXIS:

inspection (visiting the environment in which Ag_2 is working to ascertain if everything is as expected);

internal inspection (inspecting some inside agent data to check its reasoning, agenda, plan library, plan, etc.);

surveillance (by sample) observing Ag_2's activity and partial results, and the environment for avoiding damages;

detecting analyzing some traces of Ag_2's activity in order to (abductively) check whether its behaviour has been correct and at what stage it is.

COMMUNICATION:

report requests ('let me know what is happening'; 'any news?');

inspective questions ('is everything as expected?' 'what are you doing?' 'is p true?').

Examples of the Intervention

Let us now show you how the *intervention* actions can be expressed in praxis and in communication.

PRAXIS:

substitution, Ag_1 performs (part of) an action previously allocated to Ag_2;

support, Ag_1 modifies the conditions of the world so that Ag_2 can successfully perform its action or damages can be prevented;

brake, Ag_1 stops Ag_2's activity (either by external obstacles or directly acting upon/in Ag_2's body or software);

tuning, Ag_1 modifies and corrects Ag_2's action (either by external obstacles or directly acting upon/in Ag_2's body or software);

repair, Ag_1 acts in order to repair damages as a result of Ag_2's action and to recover from failures.

COMMUNICATION:

alert/warning, Ag_1 alerts Ag_2 to unexpected events or possible danger;

advice, Ag_1 provides Ag_2 with some possible recipe, solution or better action just as a piece of advice (Ag_2 is free to accept or not);

instructions, Ag_1 gives instruction to Ag_2 about how to proceed (Ag_1 is specifying (partially closing) the previously 'open' delegation);

threats, Ag_1 threats Ag_2 to induce it to do what Ag_2 should do;

reminding, Ag_1 reminds Ag_2 about what it should do; *stop*, Ag_1 orders Ag_2 to stop its activity; *abort*, Ag_1 stops delegating to Ag_2.

An interesting development of this work would be to model when and why a given form of intervention (for example: to stop Ag_2) is useful or better than others; and what kind of feedback (for example: surveillance versus report) is appropriate for a given task level of trust and possible kind of intervention.

7.2.6 From Delegation Adjustment to Autonomy Adjustment

As we said, a delegation adjustment does not always produce a change in the trustee's autonomy (by limiting, restricting or, vice versa, enlarging, expanding it). The main causes of autonomy adjustment are the following:

- there is a change of Ag_2's entitlement at the meta-level (Ag_2 can refuse, negotiate, change the delegation); or it takes such an initiative even though it is not entitled (*meta-autonomy adjustment*);
- the new task is more or less *open* than the former (*realization-autonomy adjustment*);
- there is more or less control on Ag_2 (*control-dependent autonomy adjustment*);
- there is a change in the strength of delegation (*interaction-dependent autonomy adjustment*).

Each of these autonomy adjustments can be *bilateral* (realized by either the client or the contractor or both) and *bidirectional* (either augmenting or reducing the autonomy itself). These adjustments are, at least in a significative part, strictly linked with the Ag_1's trust in Ag_2 and/or with the selftrust of Ag_2 himself.

7.2.7 Adjusting Meta-Autonomy and Realization-Autonomy of the Trustee

By crossing the first two kinds of adjustment (meta-autonomy adjustment and realization-autonomy adjustment) with the delegation adjustments, we obtain the results shown in Table 7.6: rows 1–3 show the adjustments of delegation by the trustee (trustee's adjustments)

Table 7.6 How autonomy changes while adjusting delegation and help

	Meta autonomy	Autonomy of realization
Reduction of Help	*Increased*	*Equal*
Extension of Help	*Increased*	*Increased or Equal*
Critical Help	*Increased*	*Increased*
Reduction of Delegation	*Equal*	*Reduced of Equal*
Modification of Delegation	*Equal*	*Increased or Equal*
Critical Delegation	*Equal*	*Increased or Equal or Reduced*
Openess of Delegation	*Equal*	*Increased*
Closing of Delegation	*Equal*	*Reduced*

Source: Reproduced by Permission of © 2001 IEEE.

while rows 4–8 show the adjustments by the trustor (trustor's adjustments) on its own previous delegation. In particular, we can see that:

- *When there is a trustee's adjustment there is always a change of its meta-autonomy* (the trustee decides to change the trustor's delegation); while not always there is a change in its realization autonomy. For example, in the *reduction of help,* realization autonomy remains the same because the trustee realizes just a part of the delegated task (but this part was also included in the previously delegated task). In other words, the trustee does not change autonomy as for how to realize τ. Conversely in the *extension of help*, there are two possibilities: i) the trustee has more realization autonomy when the adopted plan includes some (not delegated) part which is not completely specified (thus the delegee has more discretion in its realization); ii) the trustee has the same realization autonomy if the adopted plan does not need more discretion than the delegated one. Finally, in *critical help*, given its possibility to choose new actions, there is always more realization autonomy.
- *When there is a trustor's adjustment the trustee's meta-autonomy never changes* (in fact, the trustor itself takes the initiative to modify the delegation). As for the trustee's realization autonomy we can say that: in the *reduction of delegation* case, Ag_2's autonomy of execution (if its discretionary power is reduced with the new delegation) is reduced or it remains unchanged (suppose that the old task was completely specified in all details). In the *extension of delegation* case, either the autonomy of realization increases (if the new task presupposes some action – not included in the old one – with a certain degree of openness) or it remains unchanged (if this new task was completely specified in all details). In the *critical delegation* case, the autonomy of realization of the trustee increases or not depending respectively on whether the new actions are more or less open than the old ones. In the *openness of delegation* case, the autonomy of realization of the trustee always increases (openness is in fact a factor that increases the discretion of the trustee). Vice versa, in the case of *closing of delegation*, the trustee's autonomy of realization is always reduced.

7.2.8 Adjusting Autonomy by Modyfing Control

As already observed a very important dimension of autonomy is the control activity of the adopted/delegated task. Given that control is composed of feedback plus intervention, adjusting it means having to adjust (at least one of) its components.

Adjusting the Frequency of the Feedback

We showed in Section 7.1.2 as the *frequency of the feedback on the task* can be:

- *purely temporal* (when the monitoring (*Mon*) or the reporting (*Rep*) is independent of the structure of the activities in the task);
- *linked with the task phases* (when the activities of the task are divided in phases and the *Mon* or the *Rep* is connected with them).

Trustor and trustee can adjust the frequency of their feedback activity in three main ways:

- by *changing the temporal intervals* fixed at the beginning of the task delegation or task adoption (when the *Mon*/*Rep* is purely temporal);
- by *changing the task phases* in which the *Mon*/*Rep* is realized with respect to those fixed at the beginning of the task delegation;
- by *moving from* the purely temporal *Mon*/*Rep* to the task phases *Mon*/*Rep* (or vice versa).

Adjusting the Frequency and Kind of Intervention

As explained in Section 7.1.2, the intervention is strictly connected with the presence of the *Mon*/*Rep* on the task, even if, in principle, both the intervention and the *Mon*/*Rep* could be independently realized. In addition, the occurrence of intervention and *Mon*/*Rep* are also correlated. More precisely, the intervention can occur:

1) *never*;
2) *just sometimes* (during some phase or at specified times, a special case of this is at the end of the task);
3) *at any phase or at any time (depending on the necessity)*.

The *adjustment of the frequency of intervention* by the trustor is an important case of adjustment of the trustee's autonomy. Suppose that at the beginning there is an agreement about the fact that the established frequency of intervention is *never*, and suppose that the trustor intervenes once or twice during the trustee's task realization: the trustee's autonomy has been reduced. In general, a trustee is more autonomous if the frequency of the trustor's intervention is low. So the adjustments by the trustor in this direction (low frequency of interventions) produce an increase of trustee's autonomy. If the trustor adjusts the possible kind of intervention established at the beginning of delegation this might increase or reduce the trustee's autonomy depending on this adjustment.

7.2.9 When to Adjust the Autonomy of the Agents

We will examine in this section the general principles (criteria) for adjusting (restricting or expanding) the trustee's autonomy by both the trustor/client/delegator, and the trustee/contractor/delegee.

Table 7.7 Reducing the trustee's autonomy

WHEN (classes of reasons):

- Ag_1 believes that Ag_2 is not doing (in time) what Ag_1 has delegated to it; and/or
- Ag_1 believes that Ag_2 is working badly and makes mistakes (because of lack of competence, knowledge, control, etc.); and/or
- Ag_1 believes that there are unforeseen events, external dangers and obstacles that perhaps Ag_2 is not able to deal with; and/or
- Ag_1 believes that Ag_2 is going beyond its role or task, and Ag_1 is not happy about this (because of lack of trust or of conflict of power)[27]

THEN (reduction of autonomy) Ag_1 will reconsider its delegation to Ag_2, and Ag_2's level of autonomy in order to reduce it by either specifying the plan (task) or by introducing additional control, or constraining the interaction (strong delegation), etc.

Table 7.8 When to expand the trustee's autonomy

WHEN (classes of reasons):

- Ag_1 believes that Ag_2 is doing or can do better than previously expected (predicted); and/or
- Ag_1 believes that the external conditions are more favorable than expected; and/or
- Ag_1 believes that Ag_2 is working badly and makes mistakes (because of lack of flexibility, or because of too much control, etc.) and/or
- Ag_1 believes that Ag_2 can do more than previously assigned, or can find its own situated way of solving the problem

THEN (expansion of autonomy) Ag_1 will change the delegation to Ag_2, and Ag_2's level of autonomy in order to expand it by either letting the plan (task) less specified or reducing control or making the interaction weaker, etc.

Adjusting the Autonomy of Trustee

In this preliminary identification of reasons for autonomy adjustment, we prefer a more qualitative and simplified view, not necessarily related with a probabilistic framework like the one we will use in the following. Of course, to be more precise, one should specify that what changes is the subjective probability assigned to those events (beliefs). For example, at the time of the delegation, Ag_1 believed that the probability of Ag_2's mistakes was *pm* (and this expectation was compatible with the decision of delegating a given degree of autonomy), while Ag_1 realizes that this probability has changed (higher or lower than expected).

Let us simplify the issue in Table 7.7, Table 7.8, Table 7.9, and Table 7.10.

Trust as the Cognitive Basis for Adjusting Autonomy

What we have just seen (principles and reasons for bilateral delegation and autonomy adjustment) can be considered from the main perspective of trust.

[27] Notice that in all those cases the trustee's expectations on which trust and reliance were based are disappointed.

Table 7.9 When to limit one's own autonomy

Let us now consider some (*collaborative*) reasons of adjustment on the delegated agent's side.
WHEN (classes of reasons):

- Ag_2 comes to believe that it is not able to do the complete task (level of self-confidence); and/or
- Ag_2 comes to believe that there are unforeseen events, external dangers and obstacles that are difficult to deal with

THEN (reduction of autonomy)
Ag_2 will reconsider the received delegation (for example providing sub-help and doing less than delegated) and its level of autonomy in order to reduce it by either asking for some specification of the plan (task) or for the introduction of additional control (example: 'give me instructions, orders; monitor, help, or substitute me').

Table 7.10 When to expand one's own autonomy

WHEN (classes of reasons):

- Ag_2 gets to a grounded belief that it is able or in condition of doing more or providing a better solution for the delegated goal (within Ag_1's plan, or also with regard to Ag_1's other desires and interests), and
- it is not forbidden or it is (explicitly or implicitly) permitted by Ag_1 that Ag_2 takes such a collaborative initiative, and/or
- Ag_2 believes that Ag_1 will accept and enjoy its initiative (because convenience largely exceeds surprise or distress)

THEN (expansion of autonomy)
Ag_2 will reconsider the received delegation and level of autonomy in order to go beyond those limits by directly providing, for example, over-help or critical-help (doing more and better).

(When the 2nd and 3rd conditions above are not realized, Ag_2 could take the initiative of communicating by offering the new solution or asking for a permission, and in fact for re-negotiating the delegation).

Trust, being the mental ground and counterpart of delegation, plays the main role in adjustment: limiting autonomy is usually due to a trust crisis (Section 6.7), while expanding autonomy is usually due to an increased trust.

In Section 3.4 and Section 3.6 we have shown how *changing the credibility degree of some beliefs should change the final choice* about the delegation (and the same holds for the utilities and for the control). Resuming and concluding: *The trustor's adjustment reflects a modification in her mental ingredients*. More precisely, the trustor either *updates or revises* their delegation beliefs and goals, i.e.:

a) either they revise their *core trust beliefs* about Ag_2 (the latter's goals, capabilities, opportunities, willingness);

b) or they revise their *reliance beliefs* about: i) their dependence on Ag_2, or ii) their preference to delegate to Ag_2 rather than to do the job themselves, or to delegate to Ag_3 (a third agent) or to renounce the goal;
c) or they changes their risk *policy* and is more or less likely to accept the estimated risk (this means that the trustor changes either their set of utilities ($U(Ag_1,t_0)$) or their set of thresholds. In other words, either Ag_1*'s* trust on Ag_2 is the same but their preferences have changed (including their attitude towards risk), or Ag_1 has changed their evaluations and predictions about relying on Ag_2.

The modifications showed in the cases (a, b, c) might produce delegation adjustments but also they could suggest to the trustor the introduction of control actions (either as monitoring or as intervention). So the relationships among trust, control, autonomy, and delegation are very complex and not so simple to predict: also the trustor and trustee personalities can play a relevant role in these relationships.

7.3 Conclusions

As already shown, the relationships among *Trust*, *Control* and *Autonomy* are very complex and interesting. *Autonomy* is very useful in collaboration and even necessary in several cases but it is also risky – because of misunderstandings, disagreements and conflicts, mistakes, private utility, etc. The utility and the risk of having an autonomous collaborator can be the object of a trade-off by maintaining a high level of interactivity *during* the collaboration, by providing *both* the trustor/delegator/client and the trustee/delegee/contractor with the possibility of having initiative in interaction and help (*mixed initiative*) and of *adjusting* the kind/level of delegation and help, and the degree of autonomy run-time. This also means providing channels and protocols – on the delegator's side – for *monitoring* (reporting, observing, inspecting), and for *intervention* (instructions, guidance, helps, repair, brake); and – on the delegee's side – providing some room for discretion and practical innovation; for both client and contractor, channels and protocols are needed for communication and *re-negotiation* during the role-playing and the task execution.

Our model also provides a principled framework for adjusting autonomy on the basis of the degree of trust and of the control's level of the trustor. In particular we have shown that in order to adjust autonomy one should in fact adjust the delegation/help relationship. Thus a precise characterization of different dimensions of delegation and of goal-adoption is necessary. Moreover, we argued that adjustment is *bi-directional* (one can expand or reduce the delegee's autonomy) and is *bilateral*; not only the trustor or the delegator but also an adaptive/intelligent delegee, the trustee (the 'agent') can change or try to change its level of autonomy by modifying the received delegation or the previous level/kind of help. This initiative is an additional and important aspect of its autonomy. We showed how trust, being also the mental ground and counterpart of delegation, plays a major role in adjustment: limiting autonomy is usually due to a trust crisis, while expanding autonomy is usually due to an increased trust. Collaborative conflicts are mainly due to some disagreement about the agent's trustworthiness.

We assume that this theoretical framework can also be useful for developing principled systems.

We have outlined:

- the criteria about when and why to adjust the autonomy of an agent (for example, when one believes that the agent is not doing (in time) what it has been delegated to do and/or is working badly and makes; and/or one believes that there are unforeseen events, external dangers and obstacles that perhaps the agent is not able to deal with); and
- possible protocols of both monitoring and inspection, and of physical or communicative intervention, that are necessary for control and adjustment.

A very important dimension of such an interaction has been neglected: the normative dimension of empowerment and autonomy (entitlement, permission, prohibition, etc.) which is related to a richer and institutional relation of delegation. Also, this dimension is a matter of run-time adjustment and must be included as a necessary component when modeling several forms of interactions and organizations.

Another important issue for future works is the acceptable limits of the agent's initiative in helping. Would, for example, our personal assistant be too intrusive by taking care of our 'interests' and 'needs' beyond and even against our request (*Hyper-critical help*)? Will the user/client like such a level of autonomy or would they prefer an obedient slave without initiative? Let us leave this question unanswered as it is enough to have characterized and delimited the complex framework of such an issue.

Finally, we will leave for another book a rather important clarification for engineering: does the implementation of such a model necessarily require deliberative agents?

In fact our framework for collaboration and adjustable autonomy is presented in terms of cognitive agents, i.e. of agents who have propositional attitudes, reason about plans, solve problems, and even assume an 'intentional stance' by having a representation of the mind of the other agent. This can be exemplified via some kind of BDI agent, but in fact it is more general (it does not only apply to a specific kind of architecture). We present our framework in a cognitive perspective because we want to cover the higher levels of autonomy,[28] and also the interaction between a human user and a robot or a software agent, or between humans. However, the basic ontology and claims of the model could also be applied to non-cognitive, merely rule-based agents.

Obviously, a cognitive agent (say a human) can delegate in a weak or mild sense a merely rule-based entity. Strong delegation based on mutual understanding and agreement cannot be used, but it can be emulated. The delegated device could have interaction protocols and reactive rules such that if the user (or another agent) asks to do something – given certain conditions – it will do that action. This is the procedural emulation of a true 'goal adoption'.

Our notions could in fact be just embedded by the designer in the rules and protocols of those agents, making their behavior correspond functionally to delegation or adoption, without the 'mental' (internal and explicit) goal of delegating or of helping. One could, for example, have fixed rules of over-help like the following ones:

[28] In our view, to neglect or reduce the mental characterization of delegation (allocation of tasks) and adoption (to help another agent to achieve its own goals) means, on the one hand, to lose a set of possible interesting kinds and levels of reliance and help, and, on the other hand, not to completely satisfy the needs and the nature of human interaction that is strongly based on these categories of cooperation.

<if Ag_x asks for departure time, provide departure time & gate> (over-answering);

<if Ag_x asks for action α that has result r & not able to perform α & able to perform α' with the same result r, then perform α'>

The previous behaviors are in fact a kind of *over-help* although the performing agent does not conceive any help (the real adopter is the programmer writing such a rule).

The same remains true in the case of a rule-based delegated agent: the agent could have simple *rules* for abstaining from doing everything itself α ' while inducing – via some protocols – the needed action in another agent, or for abstaining from doing by itself α' when receiving information (communication protocol; observation) about another agent already doing the needed action.

In sum, several basic phenomena and issues (of delegating, adopting, monitoring, intervening, changing delegation, etc.) are held and recognized also by non-cognitive agents and can be incorporated in a procedural emulation of a really social interaction.

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. (1995) Guaranties for Autonomy in Cognitive Agent Architecture. In M. Wooldridge and N. Jennings (eds.) *Intelligent Agents*. Springer. LNAI 890, 56–70.

Castelfranchi, C. (1996) Commitment: from intentions to groups and organizations. In *Proceedings of ICMAS'96*, S. Francisco, June, AAAI-MIT Press.

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

Castelfranchi, C. (2000) Founding Agent's Autonomy on Dependence Theory, in *Proceedings of ECAI'00*, Berlin, August 2000.

Castelfranchi, C. and 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 August, pp. 31–37.

Castelfranchi, C. and Falcone, R. (1997) From task delegation to role delegation, in M. Lenzerini (Editor), *AI*IA97: Advances in Artificial Intelligence*, Lecture Notes in Artificial Intelligence, 1321. Springer-Verlag pp. 278–289.

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

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

Castelfranchi, C. and Falcone, R. (2000) Social trust: a cognitive approach, in *Deception, Fraud and Trust in Virtual Societies* by Castelfranchi, C. and Yao-Hua, Tan (eds.), Kluwer Academic Publisher.

Castelfranchi, C. and Falcone, R. (2000) Trust and control: a dialectic link, *Applied Artificial Intelligence Journal*, Special Issue on 'Trust in Agents' Part 1, Castelfranchi, C., Falcone, R., Firozabadi, B., Tan, Y. (Editors), Taylor & Francis 14 (8).

Castelfranchi, C. and Falcone, R. (2000) Conflicts within and for collaboration, in C. Tessier, L. Chaudron and H. J. Muller (eds.) *Conflicting Agents: Conflict Management in Multi Agent Systems*, Kluwer Academic Publishers, pp. 33–61.

Cohen, Ph. and Levesque, H. (1987) 'Rational Interaction as the Basis for Communication'. Technical Report, N°89, CSLI, Stanford.

Durkheim, E. (1997) *The Division of Labor in Society*, New York: The Free Press.

Falcone, R, (2001) Autonomy: theory, dimensions and regulation, in C. Castelfranchi and Y. Lesperance (eds.) *Intelligent Agents VII, Agent Theories Architectures and Languages*, Springer, pp. 346–348.

Falcone, R. and 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. and Castelfranchi, C. (2002) Issues of trust and control on agent autonomy, *Connection Science*, special issue on 'Agent Autonomy in Groups', 14 (4): 249–263.

Ferguson, G. and Allen, J. (1998) TRIPS: An Integrated Intelligent Problem-Solving Assistant, *Proc. National Conference AI (AAAI-98), AAAI Press*, Menlo Park, Calif.

Haddadi, A. (1996) *Communication and Cooperation in Agent Systems*, The Springer Press.

Hearst, M. (editor) (1999) Mixed-initiative interaction - Trends & Controversies, IEEE Intelligent Systems, September/October 1999.

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

Hexmoor, H. (editor) (2000) Special Issue on autonomy control software, *Journal of Experimental & Theoretical Artificial Intelligence*, 12 (2), April-June 2000.

Huhns, M. and Singh, M. (1997) Agents and multiagent systems: themes, approaches, and challenge, in *Reading in Agents* (Huhns, M. and Singh, M. Editors) Morgan Kaufmann Publishers, Inc., San Francisco, California.

Jones, A. J. I. and Sergot, M. (1996) A formal characterisation of institutionalised power. *Journal of the Interest Group in Pure and Applied Logics*, 4 (3).

Martin, C. E. and Barber, K. S. (1996) Multiple, Simultaneous Autonomy Levels for Agent-based Systems, in *Proc. Fourth International Conference on Control, Automation, Robotics, and Vision*, Westing Stamford, Singapore, pp. 1318–1322.

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

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

Pollack, M. (1990) Plans as complex mental attitudes in Cohen, P. R., Morgan, J. and Pollack, M. E. (eds.), *Intentions in Communication*, MIT Press, USA, pp. 77–103.

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

Smith, R. G. (1980) The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*.

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

van der Vecht, B. (2008) Adjustable autonomy, controlling influences on decision making, *Doctoral Thesis*, Utrecht University, The Netherlands.