2

Socio-Cognitive Model of Trust: Basic Ingredients

Trust means different things, but they are systematically related one with the other. In particular we analyze *three crucial concepts* that have been recognized and distinguished in the scientific literature. Trust is:

- A mere *mental attitude* (prediction and evaluation) towards an other agent; a simple *(pre)disposition*.[1]
  This mental attitude is in fact an opinion, a judgment, a preventive evaluation about specific and relevant 'virtues' needed for relying on the potential trustee, but that might remain separated from the actual exercise of trust.
- A *decision* to rely upon the other, i.e. an *intention* to delegate and trust, which makes the trustor 'vulnerable' (Mayer *et al.*, 1995).
  This is again a mental attitude, but it is the result of a complex comparison and match among the preventive evaluations of the different potential trustees, about the risks and the costs, and about the applicability of these evaluations to the actual environments and context.
- A *behavior*, i.e. the intentional *act* of (en)trusting, and the consequent overt and practical *relation* between the trustor and the trustee.
  This is the consequent act, behavior, of the trustor, generally coordinated and coherent with the previous decision; and the public 'announcement' and social relation.

Trust is in general all these things together.

In order to understand this concept and its real meaning we have to analyze it, and show the complex relations that exist between different terms and constituents.

[1] Or even just a feeling, an affective disposition where those 'beliefs' are just implicit (see Chapter 5).

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

2.1 A Five-Part Relation and a Layered Model

In our view trust is a *relational* construct between:

- An agent ***X*** (*the trustor*) (we will name Xania, a woman). This agent X is necessarily an 'intentional entity' (see intentional stance (Dennett, 1989)); let's say a 'cognitive agent', that is, an agent with mental ingredients (beliefs, goals, intentions, and so on). Trust must be (also) be extended in cognitive terms, as X's specific mental attitudes towards other agents and about given tasks.
- An addressee ***Y*** (*the trustee*) that is an 'agent' in the broader sense of this term (Castelfranchi, 1998; 2000a): an entity able to cause some effect (outcome) in the world; the outcome X is waiting for. When Y is an intentional agent: we will name it Yody, and he is a man.
- Such a 'causal' process (*the act, or performance*) and its result; that is, an act α of Y possibly producing the outcome p; which is positive or desirable because it includes (or corresponds to) the content of a goal of X ($Goal_X(g)=g_X$), the specific goal for which X is trusting Y. We call this act: Y's *task:* τ. τ is the couple (α, p), with g included in p or in some cases $p \equiv g$.
- $g_X = Goal_X(\boldsymbol{g})$ is in fact a crucial element of the trust concept and relation, frequently omitted.
- A context (***C***) or situation or environment where X takes into account Y (thus affecting X's evaluation and choice of Y) and/or where Y is supposed to act and to realize the task (thus affecting the possibility of success).

In other words, in our model trust is a five-part relation (at least):

$$TRUST(X\ Y\ C\ \tau\ g_X) \tag{2.1}$$

that can be read as X trusts (*in*) Y in context C *for* performing action α (executing task τ) and realizing the result p (that includes or corresponds to her goal $Goal_X(g)=g_X$).

A deep analysis of each component is needed, and a theory of their relationships and variations.

2.1.1 A Layered Notion

As we said, we consider and analyze trust as a composed and *'layered' notion*, where the various more or less complex meanings are not just in a 'family resemblance' relation, but are embedded one into the other, and it is important the explicit theory of the relations between those layers and of the *transition* from one to the other; since the most simple form can be there without the richer one, but not vice versa.

As we said, in our theory there is a double link between these forms of trust (from the dispositional one to the active one): a conceptual link and a process/causal link (see Section 1.4).

Trust as a *layered notion* (see Figure 2.1) means:

- in its basic sense just a mental (cognitive and affective[2]) *attitude* and *disposition* towards Y (*beliefs*: *evaluations* and *expectations*); this already is a social *relation*;

[2] We will put aside here the affective nature of trust; trust as an intuitive disposition, as a feeling. Here we consider trust as an explicit (and usually grounded) judgment. We dedicate the entirety of Chapter 5 to affective trust dispositions

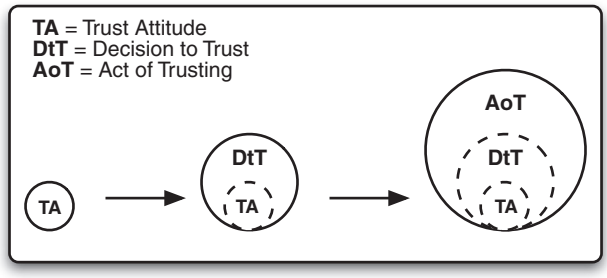

Figure 2.1 Trust stages and layers

- in its richer use, a *decision* and *intention* based on that disposition;
- and then the *act* of *relying* upon *Y*'s expected behavior;
- and the consequent overt social *interaction* and *relation* with *Y*.

The trust attitude/disposition is a determinant and precursor of the decision, which is a determinant and a precondition of the action; however, the disposition and the intention remain as necessary *mental constituents* during the next steps.

Let us introduce an example to better clarify the different concepts: Xania is the new director of a firm; Mary and Yody are employees, possible candidates for the role of Xania's secretary. Now, consider these three different aspects (and components) of Xania's trust towards them:

1) Xania evaluates Mary on the basis of her personal experience with Mary, if any, of Mary's CV and of what the others say about Mary; on such a basis she forms her own opinion and *evaluation* about Mary: how much she considers Mary trustworthy as a secretary? Will this trust be enough for choosing Mary and deciding to bet on her?
2) Now Xania has also considered Yody's CV, reputation, etc.; she knows that there are no other candidates and *decides* to count on Yody as her secretary; i.e. she has the *intention* of delegating to Yody this job/task when it comes to reorganizing her staff.
3) One week later, in fact, Xania nominates Yody as her secretary and uses him, thus actually she trusts him for this job: the trust relationship is established.

In situation (1) Xania has to formulate some kind of trust that is in fact an *evaluation* about Mary as a secretary. This trust might be sufficient or not in that situation. Since she knows that there is also another candidate she cannot decide to choose Mary (she has to wait for this) but she can just express a mere evaluation on her, and perhaps this evaluation is not good enough for being chosen (even if it has been made before next candidate's evaluation).

In situation (2) Xania arrives at a decision that is based on trust and is the decision to trust. It is also possible that – in Xania's opinion – neither Mary nor Yody are suitable persons with respect to the secretary's role and she might take the decision not to trust them, and to consider different hypotheses: i) to send a new call for a secretary or ii) to choose the least

and impulses. Consider however, that those two 'forms' of trust can coexist: there may be an explicit evaluation joined with (eliciting or based on) a feeling (Castelfranchi, 2000b).

worst between them for a brief period, looking for new candidates at the same time; or iii) to manage for a period without any secretary; and so on.

In situation (3) Xania expresses her trust through an (official) *delegation*, in fact an act of communication (or more in general, through observable external behavior). In general, this kind of relationship does not necessarily have to be official, it is simply known to the two agents (as we will see in the following, in special cases, it could also be unknown to the trustee).

In the case of autonomous agents who have no external constraints conditioning their freedom of making a reliance,[3] we can say that a sufficient value[4] of core trust is a necessary but not sufficient condition for a positive decision to trust (see Chapter 3); vice versa, a freely chosen delegation (Castelfranchi and Falcone, 1998) implies the decision to trust, and the decision to trust implies a sufficient value of core trust. We will specify which beliefs and which goals characterize the trust of the *Trustor* (X) in another agent (the *Trustee, Y*) about Y's behavior/action relevant for a given result p corresponding to (or included in) the goal of X, g_X. Given the overlap between trust and (mental) reliance/delegation, we need also to clarify their relationships.

Before starting the analysis of the basic kernel, let us analyze the relationship between p and g_X.

2.1.2 Goal State and Side Effects

As we said: $g_X \subseteq p$ (the set of the results of Y's action α contains not only X's goal but also a set of side effects). In other words, sometimes the achievement of the goal g_X is conditioned (given the specific trustee and his own features, the contextual situation, and so on) by the realization of other results in the world that could compromise other interests or goals of X.

I can trust my dentist to be able to solve my problem with my tooth, but – while going there – I know (expect) that although he will also be careful and honest I will feel some pain and will spend a lot of money. Or, I know that if I trust John – which is actually reasonable and enough for what I need –he will become too familiar, he will take liberties with me, and maybe I won't like this.

So the analysis and the knowledge of these side effects are really relevant when deciding upon trusting agents, and often a merely qualitative analysis is not sufficient (see Chapter 3).

2.2 Trust as Mental Attitude: a Belief-Based and Goal-Based Model

Our main claim is that: *only a cognitive agent can trust another agent; only an agent endowed with goals and beliefs.*

First, one trusts another only relative to a goal, i.e. for something s/he wants to achieve, that s/he desires or needs. If I don't potentially have goals, I cannot really decide, nor care

[3] Absence of external constraints is an ideal condition: in fact, also in case of autonomous agents some constraints are always present. In the case of the previous example, a constraint is given from the impossibility to evaluate all the potential candidates available in the world.

[4] In fact, trust is also a quantitative notion: it is constituted by different ingredients to which it is possible/necessary to attribute a value.

about something ('welfare'): *I cannot subjectively trust somebody*. These goals could also be maintenance goals (of a situation, a state), not necessarily achievement goals.

Second, trust itself basically *consists of* (implicit or explicit) beliefs.

The root of trust is *a mental state*, a complex *mental attitude* of an agent X towards another agent Y, in context C, about the behavior/action α relevant for the result (goal) g_X.

Since Y's action is useful to X, and X is relying on it, this means that X is *delegating* to Y some action/goal in her own mental plan. This is the strict relation between trust and reliance and delegation: *Trust is the mental counter-part of reliance and delegation.*[5]

This mental attitude is based on and consists of *beliefs* (or in general of doxastic representations[6]), about the trustee and his behavior. And in fact they may be wrong. X can have false (although well grounded and subjectively justified; rational) beliefs about Y's qualities, skills, and behavior. In short, the main beliefs are:

i) X believes that Y is able and well disposed (willing) to do the needed action;
ii) X believes that in fact Y will appropriately do the action, as she wishes;
iii) X believes that Y is not dangerous; therefore she will be safe in the relation with Y, and can make herself less defended and more vulnerable.

The first (and the third) family of beliefs is '*evaluations*' about Y: to trust Y means to have a good evaluation of him. Trust implies some appraisal.

The second (and the third) family of beliefs is '*expectations*', that is (quite firm) predictions about Y's behavior, relevant for X's goal: X both wishes and forecasts a given action α of Y, and excludes bad actions; she feels safe.

The basic nucleus of trust – as a mental disposition towards Y – is a positive expectation based on a positive evaluation; plus the idea that X might need Y's action.

Let us carefully consider these various roles of the beliefs in a trust mental state, and these two facets of trust: as valuation, as expectation.

2.2.1 Trust as Positive Evaluation

An explicit positive evaluation is a judgment, a belief about the goodness of Y for the achievement of a certain goal. 'Y is good' actually means 'Y is good for...' (Miceli & Castelfranchi, 2000). Sometimes we do not specify *for* what Y is good, just because it is included in the very concept of Y ('This is a good knife/pen/car/... mechanic/doctor/father/...'), or because it is clear in the context of the evaluation ('On which mountain we can do some climbing?' 'That mountain is good!'). These are direct explicit evaluations: 'Y is good, is OK, is apt, able, useful', and so on.

The abstract content of these predicates is that: *given the goal g_X, Y is able to realize it or to help X in realizing it* (a tool). In other words, Y has the *power of* realizing g_X.[7]

[5] Given this strict relation and the foundational role of reliance and delegation (see Section 2.6) we need to define delegation and its levels; and to clarify also differences between reliance, delegation, and trust.

[6] See Section 2.2.8 on Trust and Acceptance and Chapter 1, note 21.

[7] The evaluation of Y is about the *internal powers* for g_X (internal resources, capabilities, competences, intelligence, willingness, etc.), but for relying on Y for g_X external conditions and the control on the external conditions might also be necessary: Y may also have (or not) the *external powers*, the external conditions and resources for realizing g_X; Y

However, trust is not only a belief that Y 'is good for τ', it is also a set of beliefs about the needed qualities or 'virtues' of Y. Trust, as evaluation, is in fact a model of Y's *qualities* and *defects* (which define his *trustworthiness dimensions*).

As for any kind of explicit evaluation, with trust we are not satisfied by the mere belief that Y is OK, is 'good for', has the power to achieve goal g_X, to execute the action/task α/τ (delegated to him). We try to *understand* why Y is good for this (while Z is not); to have a *theory* of what makes Y *able*.[8] In other terms, we try to know which kind of characteristics are useful or required for effectively performing τ or achieving g_X. And many of them are just hidden, internal (and mental): *kripta* (Section 2.2.7).

Qualities and Standards

This applies in general to the Evaluation theory (Miceli and Castelfranchi, 2000).

Given that Y is 'good for' τ (for example, this knife Y is good for cutting the bread), to which features of Y should this be ascribed? In the case of the knife: To its plastic or wooden handle? To its sharpening? To its color? To its being serrate and long? And so on. In several cases this implies a true *causal model* (although naive and commonsensical) of τ, of Y, and of what effects it produces. In the example of the knife, it is clear that the plastic handle or the color are irrelevant, while a good knife for bread needs to be long, serrate, and sharpened.

Those features (F) to whom the 'goodness of Y for...' is ascribed are Y's *qualities* (Q). *Defects* (D) are those features of Y's to which is attributed the lack of power, the inadequacy or dangerousness of Y for α/τ.

Notice that a feature of Y that is a 'quality' relative to task τ, can be a 'defect' relative to another task τ', and vice versa (see Figure 2.2).

Let us also take note of how this theory of 'qualities' and 'defects' is cognitive and pragmatically very crucial and effective. In fact, while buying a new knife in a store we are not allowed to have with us a piece of bread and experimentally evaluate which knife is 'good for' it. Thus, how can we choose a 'good' knife without trying it? Just because we have a theory of the needed qualities, of what makes a knife a good knife for bread. We just look for these characteristics; we compare that knife with the 'standards' for a good bread-knife.

Standards are in fact just *qualities generalized* to the class of the object; the ideal properties that such a kind of object O *must* possess in order to be good as an O.

Qualities (and standards) are necessary not only for *recognizing* and choosing a good instance of object O; that is they are *signs* of its being good, reliable for τ (trustworthy) (see Section 2.2.7 on Signs); but they:

- are fundamental also for generalizing trust from one task to another (are for τ needed more or less the same qualities than for τ'?), or from one agent to another: 'has Z the relevant qualities remarked in Y?' (see Chapter 6); and thus they;
- are also fundamental for predictions based on general models and classes.

---

is both 'able' and 'in condition' for realizing g_X. For example, in J. J. Meyer's logic (Meyer, 1992) Y 'CanDo' when both Y is 'Able' and 'In Condition'.

[8] While we put the other dimension (the evaluation of the conditions and external resources) in the 'environmental trust' (see Section 2.10).

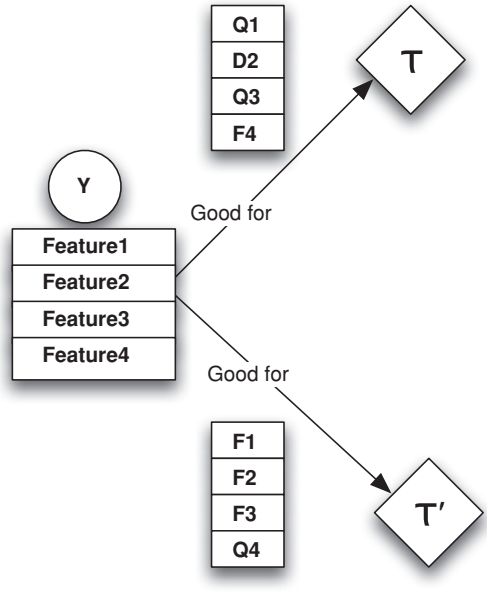

Figure 2.2 From Features (F) to Qualities (Q) through Evaluation: in particular (Feature 1 and Feature 3) are evaluated as good for the task τ while (Feature 4) is good for the task τ'

X trusts *Y* – that is, she has a positive evaluation of him for τ (and she generates positive expectations of him), also because she ascribes to him certain *qualities*.

Then, we can say that if (see Figure 2.3):

SetQ is the set of qualities (powers) needed for α/τ, then for p (g_X); in case *Y* possesses all the elements in *SetQ* then *X* could trust *Y* for achieving g_X.

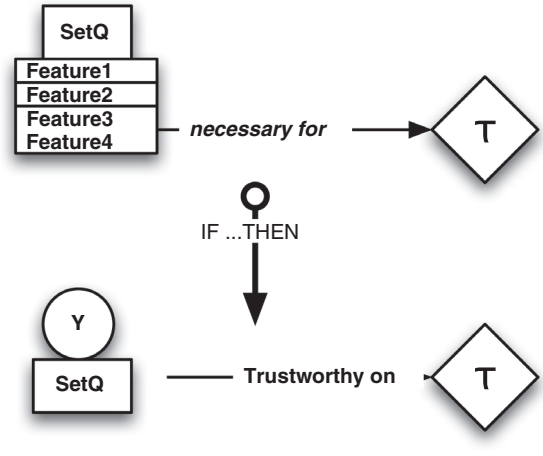

Figure 2.3 Given a Set of Abstract Qualities for a task, how to select the right Agent for that task

For any *Y*, the more the elements in *SetQ* are owned by *Y* the more trustworthy is *Y* for α/τ.

So we can say that trust is constituted of what we call 'implicit and indirect evaluations', that is of specific features (like 'sharpening; long; serrate', or for a person like 'tall, bold, agile, intelligent, ...') which only apparently are just *descriptive* predicates and usually are *evaluative* predicates. These are the specific and analytic beliefs on which the global positive or negative judgment is grounded.

Reason-based trust is in fact a *theory* about *Y*, and about his qualities and defects that make him more or less apt and reliable for α/τ; on this basis we make predictions and build expectations about *Y*'s performance and its outcome; and we explain the reasons of success or failure; like for any *theory*.

This also means that the feedback on trust due to the observed outcome of *Y*'s performance (see Chapter 6), not only can be attributed to *Y*, that is to internal factors (and not to external accidents or conditions), but in some cases can be specifically ascribed to a sub-feature of this causal model of *Y*'s *power on* τ. And one can revise this specific piece of evaluation and trust: '*Y* is not so sharp as assumed'; '*Y* is not so agile as supposed', and so on.

Trust and Powers

An evaluation is a judgment about the possible powers of *Y*. In our abstract vocabulary *Y can* (has the *power of*) perform a given action or realize a given goal. It is relative to this that *Y* is *good for*. Correspondently, *qualities* are just *powers of* (to be *strong*, to be *smart*) or conditions for *powers of* (to be *entitled*, to be *prepared*).

This is about not only personal powers (be strong, intelligent, skilled, etc.) but also social powers: be influent, prestigious, being a friend of *Z* (and thus be able to obtain from *Z* something), being sexually appealing, etc.; and also *institutional powers*, the new capacities that a given institutional role gives to its players: like nominating somebody, or proclaiming, or officially signing, or to marry, etc.[9]

So there is a complex relation between powers and trust. On the one side powers – and in particular perceived and appreciated powers – make us trustworthy for the others (they can need us and wish to rely on us), and this is why our *social image* and the *signals* that we give of our qualities are so important. As we will see, trust (being trusted in a given community, and in particular being better evaluated than others) is a *capital* (see Chapter 10).

On the other side, this is a capital precisely because it increases our powers; provides us new powers. Since we (being perceived as trustworthy) are appreciated and demanded, we can enter in exchange relations, we can receive cooperation, we can influence other people, etc. This means that we greatly enlarge our *practical powers* (the possibility to achieve – thanks to the other's cooperation – our goals) and *social powers*: powers on the others. Also for being invested by institutional roles (and receive new special powers) we need some competence and reliability, or even better we need some trust from the others on this. In sum, *powers give trustworthiness and trust from the others; and trust gives powers.*

[9] See (Searle, 1995); as for the relationships between personal and institutional powers, see (Castelfranchi, 2003).

The Recursive Nature of Trust as Mental Representation

One might claim that in a strict sense trust just is or implies the positive *expectation*, which in its turn is based on positive evaluations; but that (i) *the evaluations are not trust or* (ii) *they are not implied by trust* and *they are not strictly necessary*: One might trust Y without specific and explicit beliefs on his skills or reliability, just on the base of experience and learning.[10]

However, this is not fully true.

As for (ii), Y's trustworthiness or the trust *in* Y always imply some (perhaps vague) attribution of internal qualities (such that the task could be achieved/realized); some 'kripta' which makes Y apt and reliable (Section 2.2.5).

As for (i), the evaluations of Y, when used as the bases and reasons for a trust expectation and decision, are *subsumed* and *rewritten* as aspects and forms of 'trust'. Given that on the basis of those features (*qualities*) X trusts Y, then X trusts those features of Y. 'I trust (in) his competence', 'I trust (in) his benevolence', 'I trust (in) his morality', 'I trust (in) his punctuality', and so on.

Let us ground this on the general theory of goals and actions. Given her goal g_X (X's aim or end) X searches for some possible action α (usually of X) able to achieve g_X (where g_X is included in the outcomes p, the post-conditions of α). Given an action α useful for g_X, this action in order to be successfully performed requires some condition C to be true. If C holds the subject can perform α; if C does not hold it becomes the new goal ${g^1}_X$ of X, subordinated and instrumental to g_X: a sub-goal. This obviously is the essential abstract principle of planning.

Now, given that X has the goal of realizing g_X, and that she is dependent on Y and needs an action α_Y of Y, she has the sub-goal that Y successfully performs α_Y. However, certain conditions are needed for both (i) Y successfully performing α_Y; (ii) X can decide to count on this and counts on this.

Y's valuation is precisely the epistemic activity aimed at those conditions; and the same hold for X's predictions about Y doing or not α_Y. As we have underlined, evaluations are about goals (something is or isn't 'good *for*'), and predictions in trust are in fact not simple 'predictions' (beliefs about the future) but more rich 'expectations' (involving goals). Actually, since X wishes to achieve g_X through Y's action, and has the goal that Y be able, in condition, and predictable in performing α_Y, all these necessary conditions (also those 'internal' to Y) are new (sub)goals for X: X *wishes* that Y is skilled enough, competent enough, not hostile or dangerous, willing and reliable, and so on. Relative to those sub-goals she evaluates Y and has (or not) trust in Y. She trusts Y *for* being competent, *for* being persistent, etc.

So trust in Y as for action α_Y for goal g_X, (at least implicitly – but frequently explicitly) entails sub-trust supporting the broad trust about the action; in a recursive way. Since any new goal about Y might have its sub-conditions and needed sub-qualities on which X relies; thus potential sub-trusts are generated. For example, X can trust Y for being really willing to cooperate with her, because she knows and *trusts in* Y's friendship, or because she knows and *trusts in* Y's generosity and empathy, or because Y is morally obliged and she knows and *trusts in* Y's morality.

This is not only the real use of the word and the commonsense on trust; it is a logical and principled use, as we have just explained.

[10] We thank Fabio Paglieri for this observation. This is also one of Andrew Jones' criticisms to our model (Jones, 2002); see Section 2.2.3 for our reply.

2.2.2 *The 'Motivational' Side of Trust*

In our opinion, part of the core of trust is the 'prediction' component: a belief about a future event or state. However, this core element although *necessary* is not *sufficient* in our view (as for example claimed by A. Jones (A. Jones, 2006)).

Many definitions in the literature (Chapter 1) (even those mentioned by Jones) either explicitly or implicitly also contain the idea that trust is relative to some reliance upon, to some welfare and interest. In our analysis this means that the trustor has some 'goal', not only beliefs. Some author contexts precisely this point, that for us is fundamental: the other core (motivational).

Is Trust Reducible to a (Grounded) Belief, a Regularity-Based Prediction?

Certainly, one can establish a conventional, technical meaning of 'trust' far from its natural language meaning and psycho-sociological use; but in our view this is not particularly useful. It would be more heuristic to abstract from common-sense meaning and to enlighten and identify (and formalize) the basic, necessary and sufficient conceptual constituent (useful for a rich and explanatory theory of the psycho-social phenomena). In this case we think that one cannot miss the fact that when *X* trusts someone or something for something *X* is concerned, is involved; *X* cannot be neutral and indifferent about what will happen. In other words, *X* has not simply a prediction, but a full 'expectation'. In our analysis an 'expectation' is more than a simple forecast or belief about the future (based on some experience, law, regularity, or whatever). Trust in our model is composed of 'expectations' about *Y*, his behavior, and a desired result. A mere 'regularity' does not produce 'trust' (it can even produce 'fear'). Even to produce Luhman's 'confidence' (just based on regularities without choice) something more is necessary. In fact, confidence (which in fact is just a simpler form of trust) is a positive, pleasant feeling; it implies some implicit or explicit goal and its realization (for example to avoid stress, surprise, problems, anxiety). When apparently a mere prediction or perceived regularity determines a feeling or attitude of trust or confidence it is because *X* not only believes but desires that the world goes in such a regular, predictable way: see the need for 'predictability' ((Bandura, 1986) our theory of expectations, etc.). The issue is how predictions become prescriptions (Miceli and Castelfranchi, 2002) and (Castelfranchi *et al.*, 2003).

When we set up or activate protection or surveillance systems, precisely because we know that there is a high probability of rapine (ex. banks) or of aggressions, we do not have 'trust' they will rapinate or aggress us! And when we institute the firemen organization we 'expect' but we do not 'trust' that there will be fires! We 'trust' that firemen will be able to put out the fire (as desired!).

There are computer making/producing weather forecasts, but they do not have 'expectations' about the weather (although they could check whether the prediction were correct and fulfilled, in order to learn and adjust); even less likely do they have 'trust' about a sunny day (or rainy!).

Thus a mere belief (prediction) (even regularity based) is not enough. Nor it is 'necessary'. In fact there can be trust and bet/reliance on non-regular, exceptional events (perhaps an 'irrational' trust, but trust; *'This time it will be different! I'm sure, I feel so: it cannot happen again!'*).

Are Expectations and Trust Always Based on Regularities?

There can be expectation about conformity to rules, but also expectations about 'exceptions' or violations (they are not an oxymoron).

I'm worrying about being aggressed and robbed; actually I know (they said) that this has never happened here, but since I feel weak, alone, and without defense I really worry and expect some aggression.

Moreover, there can be new, unexplored circumstances and problems and usually – that's true –we are more careful, diffident; but we can also be fully convinced and trustful in our creative solutions and on some intuitive or reasoned prediction.

Previsions, expectations, and trust are not (for us) always and necessarily based on rules, norms, regularities (except to postulate that any 'inference', or association and learning, is by definition based on an explicit rule we trust.[11]

In conclusion, (regularity-based) beliefs/predictions are:

- Neither *necessary* (predictions are not necessarily regularity/rule based; moreover even when there is a regularity or rule I can expect and trust an exceptional behavior, event, like winning my first election or the lotto).
- Nor *sufficient*: the goal component (some wish, concern, practical reliance) is implied for us in any 'expectation' and a fortiori in any form of 'trust' which contains positive expectation.[12]

2.2.3 The Crucial Notion of 'Goal'

For our non-reductive theory of trust (not simply epistemic but motivational and pragmatic) it is crucial to make clear the central notion of the 'motivational' dimension: 'Goal'.[13]

Trust attitude is not just a (grounded) belief, a prediction; it is not just a subjective probability of an event, because this belief structure is motivated, is charged of value, is anchored to a goal of *X*.

In trust *X* is interested, concerned; the event is a 'favorable' one; *X*'s 'welfare' is involved. An 'expectation' is not a 'prediction' or 'forecast'; when *X* trusts somebody this implies a positive evaluation of him. Also affects and emotions can be involved, as the real basis of a trust disposition or complementary to the judgment and decision (for example, *X* will not just be surprised but she will be disappointed or even feel betrayed); but, in fact, *no affective reactions or emotions are possible without involved goals*.

[11] See also Section 9.6 on Norms.

[12] Thus the relationship between our model and A. Jones' model is not a relation of abstraction or inclusion (where Jones' core would be more abstract, pure, and contained in our definition: vice-versa the extension of our trust would be included in Jones' trust extension); but, it is a relation of partial overlapping: the common constituent being the prediction belief, the diverging constituents being the 'regularity belief' (not necessary for us) and the wish/goal component (non necessary for Jones). As for another Jones' critic to our model (that is: that we explain what it means to "trust", but not "why" we trust something/somebody), we reject this critic by modeling Y's trustworthiness (ascribed 'virtues' and 'powers), the perceived success and risk, the personal or social reasons Y's should have for behaving as expected, etc., as the very basis of the decision to trust.

[13] An interesting and remarkable reference to this component is given by Good (Good, 2000).

Under all these crucial aspects a motivational component is presupposed in *X*, and makes clear 'expectations', 'evaluations', 'concern' and 'interest', 'affects', etc.

Actually, when we examine all these phenomena, we explain them in terms of important *motives, needs, projects, desires, preferences, objectives,* and so on, for the realization of which the agent is evaluating other agents and the possibility to rely on (exploit) them. The abstract category we use *for all these motivational terms* and categories is 'goal'. However, it must be very clear what a 'Goal' is in cognitive psychology, on the basis of the cybernetic definition and of the following psychological models.

'Goal': What is This?

'Goal' is a perfect term but not in the most superficial and typical English use, where it usually refers to a *pursued external objective to be actively reached*; some step in my active plan, driving my action.[14]

The right general notion of 'goal' and of 'goal-directed or driven behavior' has been provided by cybernetics and control-theory many years ago (Rosenbleuth Wiener in 'Purposive Systems'), and has been imported in psychology and Cognitive Science in the 1960s with the TOTE model by Miller, Galanter and Pribram (Miller, Galanter and Pribram, 1960).

A 'Goal' is the mental representation that 'evaluates' the world (we evaluate the world against it); if the world does not match with, we can be activated for changing the world and realize that state; but, before this, we examine it to see if it is the case or not: is the goal self-realizing or impossible? Or: is it up to us to realize it? Do we have the appropriate actions for this? Are there the necessary conditions for acting successfully? Are there more important goals to be preferred? After all these tests the Goal may become our pursued objective. *But, it already is a 'Goal' in all the other conditions and previous steps.*

It is a 'Goal' even before we believe or know that it is realized or not; before we decide that it depends on us to (attempt to) realize it; that we can and should act. Then it can become an 'intention' (if chosen, if I have a plan for it, if I'm ready to; if I have decided to act for realizing it) and be 'pursued' (Castelfranchi and Paglieri, 2007).

In sum, a Goal is a Goal even before or without being pursued: happiness is due to goal-realization and sufferance is due to goal frustration, but not necessarily to our *active* successes or failures: we are crying because our mother has died, or happy because, without asking, doing or expecting anything, she gave us a kiss.

Given this – not vague, common sense, or reductive – notion, we can make clear that:

[14] Consider for example that one of Andrew Jones' central objections to our model of trust is precisely due to such a peculiar use of the notion of "goal" (Jones, 2002): *"While it is true to say that a goal-component of this sort is often present, this is by no means always so. For example, x might trust y to pay his (y's) taxes . . . , even though it is not a goal of x that y pays. Also, x might trust y when y asserts that p, even though x does not have it as a goal to find out whether p is the case."* (p. 229). On the contrary, the general notion of "Goal" precisely covers also those cases. Of course *X* has the goal that the other guys pay their taxes! Only in this sense he "*trusts* them as for paying their taxes". Obviously, this does not mean that *X* is personally doing something in order to produce this result in the world; she is not actively pursuing that goal. But not all our goals are or must be personally pursued. *X* is wishing, desiring, prescribing, expecting, ... that the others will pay taxes. That's why, just in case, she will be not just surprised but frustrated and upset, and will blame them, and so on. Analogously, "*x does not have it as a goal to find out whether p is the case"*, sure! However, *X* has the goal (wish, prescription, and so on) that *Y* says the truth; precisely in this sense "*X .. trusts Y when Y asserts that p".*

- In Trust as 'disposition', 'attitude', evaluation of *Y* (before deciding to delegate or not, before delegating) *Goals are not yet or necessarily pursued*: I evaluate 'potential' partners or goods or services relative to a *possible* goal of mine, or relative to a goal of mine that I have not yet decided if I want to achieve or pursue.
  Saying that *X* trusts *Y* relatively to a given (possible) goal of her, does not necessarily mean that she is actively *pursuing* a goal: on the one hand, she can be just evaluating a potential delegation; on the other hand, she can be completely passive, just waiting and expecting.
- In Trust, as decision and action, clearly that goal has become not only 'active' but 'pursued': I want to realize it. However, it is pursued in a strange way: thanks to the action of another agent; parasitically. It is *indirectly* 'pursued'; perhaps I'm doing nothing, just expecting and waiting for it to be realized thanks to *Y*'s behavior.
- When I do actively and personally pursue a Goal, this is a Goal in the reductive sense, and I trust myself (self-trust, self-confidence; feeling able and competent, etc.).
  This is the right, broad, scientific notion of 'goal' needed for the theory of Trust.

2.2.4 Trust Versus Trustworthiness

It is also fundamental to make clear the relationships between *trustworthiness* and *trust*; they are frequently mixed up; in many cases we tend to use trust (*T*) instead of trustworthiness (*TW*). *TW* is a property of *Y* (but in *relation* to a *potential* evaluator/partner *X*); while *T* is a property of *X* (but in *relation* to *Y*).

On the one hand, there is an *objective TW* of *Y* (what he is actually able and willing to do in standard conditions; his actual reliability on a more or less specific task, and so on). It is just relative to this reliability that *X*'s *T* can be *misplaced* and *X*'s evaluation (belief) can be *wrong*. On the other hand, there is a *perceived*, or evaluated, or *subjective TW* of *Y* for *X*: ${}_{X}TW_{Y}$.

Now, the relation between ${}_{X}TW_{Y}$ and *X*'s *T* in *Y*: *Trust*$(X\,Y\,\tau)$[15] is not simple. They cannot be identified. ${}_{X}TW_{Y}$ is one of the bases of *T*, but the latter cannot be reduced to the former. *T* is a direct function of ${}_{X}TW_{Y}$, but not only of it. *T* is also a function of a factor of *X*'s personality and general trustworthy disposition (see (McKnight and Chervany, 2001)). Moreover, *T* depends on the favorable or unfavorable ascription of the *plausibility* gap (lack of evidences); so the perceived ${}_{X}TW_{Y}$ is only one component of the positive evaluation (*T* attitude).

Of course, ${}_{X}TW_{Y}$ is even more insufficient for defining and determining *T*'s potential decision and *T*'s actual decision and act (see below and Chapter 3).

TW is *multidimensional*; thus ${}_{X}TW_{Y}$ is also a multidimensional evaluation and profile of *Y*, and cannot be collapsed in just one number or measure. The same holds for trust. We will in fact present a multi-dimensional model of *T* (and *TW*) (Section 2.2.6).

2.2.5 Two Main Components: Competence Versus Predictability

The (*positive*) *evaluation* of *Y* has different aspects and is about different *qualities* of *Y*. The most important dimensions for trust, that is, of *Y*'s trustworthiness, are the following ones.

[15] We consider $p{=}g_X$ and do not consider here the context *C*.

Competence

Competence is the set of qualities that makes *Y able* for τ; *Y*'s internal powers: skills, know how, expertise, knowledge, self-esteem and self-confidence,[16] and so on. When *X* trusts *Y* for τ, she ascribes to *Y* some competence.

Competence – as we claimed – cannot be put aside (like in many models and definitions; see Chapter 1) and cannot be separated from trust in *Y*'s reliability.

First of all, it is an important issue in rejecting the fully normative foundation of trust – pursued by many authors (like (Elster, 1979); (Hertzberg, 1988); (Jones, 2002)), which cannot be extended in a simple way *Y*'s skills and competences.[17]

Moreover, competente and reliability are not fully independent dimensions. Even *Y*'s cooperative (*adoptive*) attitude towards *X* may require some skill and competence. For example, *Y* may be more or less competent and able in *understanding X*'s needs or in comprehending *X*'s interests (even beyond *X*'s own understanding); it is not just a matter of 'concern' or of good will. For example, *Y*'s competence, ability, expertise, can be the basis for his self-confidence and sense of mastering, and this can be crucial in *Y*'s willingness and intention to adopt *X*'s goal, in *Y*'s persistence in this intention, which are crucial aspects of *Y*'s 'reliability'. And so on.

Predictability and Willingness

The second fundamental dimension is not about *Y*'s potential and abstract capability of doing τ, but about his actual behavior; the fact that *Y* is reliable, predictable, one can count on him; he not only *is able to do*, but *will actually do* the needed action.

Applied to a cognitive *Y*, this means that *Y* is *willing* (really has the intention to do α for g_X) and *persistent* (Castelfranchi and Falcone, 1998a), (Falcone and Castelfranchi, 2001b). Also, in this case we have to consider on the one hand, the abstract predictability and willingness of *Y* not related to other elements and, on the other hand, its relevant correlations with the specific tasks and the different trustors.

These (*Competence and Willingness*) are the prototypical components of trust as an attitude towards *Y*. They will be enriched and supported by other beliefs depending on different kinds of delegation and different kinds of agents; however, they are the real cognitive kernel of trust. As we will see later, even the goal can be varied (in negative expectation and in aversive forms of 'trust'), but not these beliefs.

Those (evaluative) beliefs are not enough; other important beliefs are necessary, especially for moving towards the decision to trust, and the intention and action of trusting.

Using Meyer, van Linder, van der Hoek *et al.*'s logics ((Meyer, 1992), (van Linder, 1996)), and introducing some 'ad hoc' predicate (like WillDo)[18] we can summarize and simplify the mental ingredients of trust as follows:

---

[16] In rational beings (which decide to act on the basis of what they believe about the possibility of achieving the goal) there is a strange paradox of power: It is not enough 'to be able to'; in order to really be able, having the power of, the agent must also believe (be aware) of having the 'power of', otherwise they will renounce, they will nor exploit their skills or resources. (Castelfranchi, 2003)

[17] Although this can be made simpler precisely by our theory of 'standards' as the needed, expected, and thus 'prescribed' qualities that a given instance of class O must possess in order to be a good or regular O. However, in any case, one has to distinguish 'normative' from 'moral'.

[18] This is a simplification. Before being a belief that "*Y* will do" this is a belief about a potential delegation: "*Y* (in case) would do" "*Y* would be able and willing to . . .", "*X* might rely on *Y*". (See Section 2.3.2).

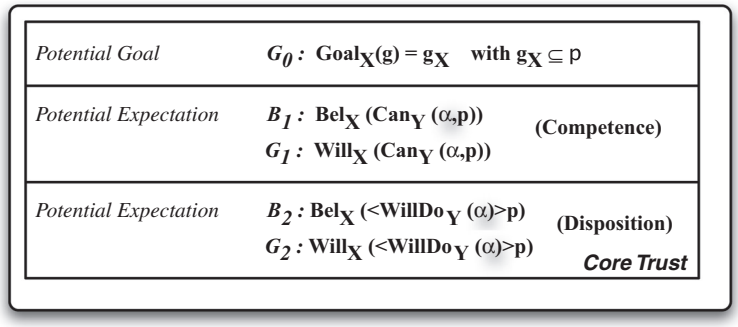

Figure 2.4 Mental Ingredients of the 'core trust'

In Figure 2.4 we simplify and summarize the kernel of trust as potential evaluation and expectation for a potential delegation.

Notice that Competence, Willingness (Predictability), and also Safety are three necessary components and *Dimensions* of trust and trustworthiness. This doesn't mean that in order to trust *Y* (and possibly and eventually to decide to trust him) *X* should necessarily have a good evaluation of *Y*'s competence and of *Y*'s willingness. As will be clear later, after introducing the degree of believing as the basis for the degree of trust, trust is not a yes/no object; only a trust decision eventually is a yes/no choice, and clearly needs some threshold. So *X*'s trust in *Y* (evaluation of trustworthiness) must be just *sufficient* (and frequently just in a comparative way) for taking a risk on him. Perhaps competence is perceived as rather low, but altogether the positive evaluation and expectation is enough. Of course there might also be a specific threshold for a given dimension: 'no less than this'; and in this case *X* must focus on this and have an explicit perception or evaluation of it.

For example, we assume that we have a threshold of *risk acceptance*: although convenient, we may refuse a choice which involves more than *a certain* risk. What we claim is just that (explicitly or implicitly) these *dimensions* about *Y*'s ability and know how, about his predictability and reliability, about their safety, are there, in the very disposition to trust and entrust *Y*.

2.2.6 Trustworthiness (and trust) as Multidimensional Evaluative Profiles

As we have seen, both while explaining the theory of *qualities*, and when analyzing the basic constituents or dimensions of trust evaluation (*competence* and *willingness*), which can be further decomposed and supported: Trustworthiness is not just a simple mono-dimensional quality. It is the result of several dimensions. We can, for example, consider two rather independent dimensions under the notion of *competence*: the *skills* or *abilities* of *Y* versus his *know how* (knowledge of recipes, techniques, plans for)[19]; and other rather independent dimensions around the *willingness* in social trust: *Y*'s *concern* and *certainty of adoption* versus his *persistence in intending* (Figure 2.5).

[19] One can have a perfect knowledge about "how to do" something, but not be able to, since one lacks the necessary skills and abilities; vice versa, one might in principle be able to successfully perform a given activity, but lack the necessary "know how": the instructions or recipes about how to do it.

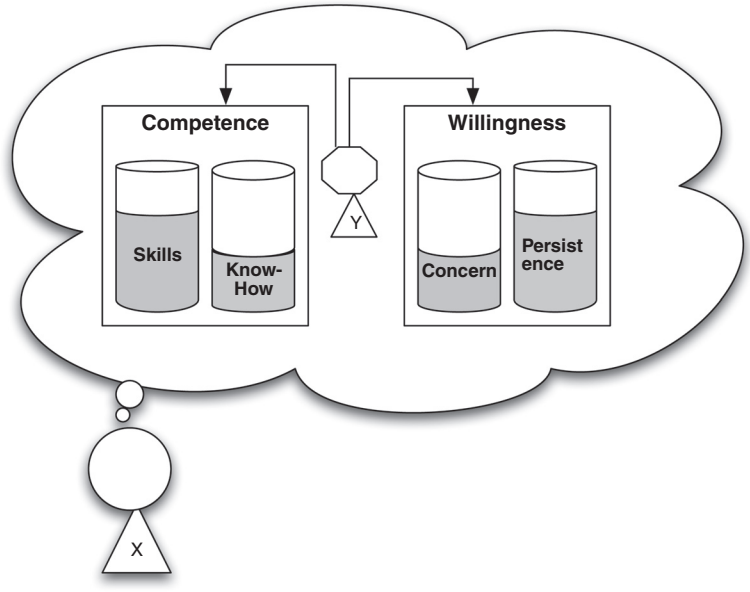

Figure 2.5 Dimensions of Trust Evaluation and their sub-components

2.2.7 The Inherently Attributional Nature of Trust

The very attitude, act, and relation of trust of *X* in *Y* (for a given performance) implies *X*'s *causal internal attribution* of the possibility of success. Trust is not simply a prediction (although it implies a prediction). In particular, it is not a prediction of a given behavior based on some observed frequency, on some estimated probability, or on some regularity and norm. It requires the grounding of such a prediction (and hope) on *an internal attribution to Y*.

This is why one trusts *in Y*, and trusts *Y*. Also the 'trust *that* (event)' (that something will happily happen) cognitively and procedurally implies a 'trust in' something or somebody on which one relies, that is, the assumption that this entity will produce the desired event.

Trusting (in) *Y* presupposes a possibly primitive and vague 'causal mental model' (Johnson-Laird, 1983) one that produces the expected result. Even when trust is non-social, when for example we decide to trust a *weak chair* by sitting on it, we assume that its material, or its mechanics, or its structure will resist under our weight. Trust presupposes at least a *design* stance (to use Dennet's terminology[20] (Dennet, 1989)).

This is the deepest meaning of the *competence* ascribed to *Y*, of the internal *power of* appropriately executing the task; this is the real meaning of the predicate *Able* used for its decomposition. It is different from the other component. In order to perform the task *Y* must be both *Able* and *in condition* but while *Able* is an internal attribution, *in condition* can be external and contextual.

[20] Although we believe that our theory of 'functioning' versus function (Miceli and Castelfranchi, 1983) is somewhat clearer.

Moreover, this is also the deep meaning of the other component: *willingness, intention, disposition, motivation* of *Y* that makes that prediction grounded.

In other words, in social trust the necessarily internal causal attribution of the trust requires an *intentional stance* towards *Y* (Dennett, 1989). *Trust is a disposition (and an affect) and a social decision only possible in agents able to assume an 'intentional stance', that is to ascribe a mind and mental stuff to the other agent*. The prediction of the other's behavior is based on some theory of mind or on some projection/ascription of *internal* factors (including moods and emotions) playing a causal role in the activation and control of the behavior.

The fact is that in humans those *qualities* are *internal* in a psychic sense; thus – by definition – they are *unobservable*. They can just be aduced by external signs. They are 'kripta' ((Bacharach and Gambetta, 2000), (Bacharach and Gambetta, 2001)); but it is precisely in 'kripta' that we trust and on 'kripta' that we count on when we 'trust' somebody; not the 'manifesta' (visible signs). Trust is a hermeneutic and semiotic act. We trust the signs only for being reliable, for informing us about the 'kripta', but *at the end it is of /on/for the 'kripta' that we trust.*

'Manifesta', signs, are not only the direct observable behaviors or markers of *Y*, but also all the other *sources* of trust (of trust beliefs) – like *experience*, *reputation*, *certificates*, *category*, *role*, and so on. All of them are direct or indirect *manifestations* of how *Y* is or 'works', of what makes him trustworthy.

Our favorite example of *weak delegation* (Section 2.6.1) makes our claim clear: I'm running to take the bus, the driver and the people cannot see me, but I see that there are people at the bus stop apparently 'waiting for the bus' (notice that this is just an mental ascription), on such a basis – attributing them with the intention to take the bus and thus the intention to stop the bus – I rely on them, I trust them for this.

While a *simple prediction* that somebody standing at the bus stop might raise their arm would be based on simple observed frequency and probability, we deny that trust in *Y* to stop the bus is just this. It is based on the interpretation of *Y* as 'waiting for the bus' and *on the ascription of the intention that* will cause the action on which one relies. This is the – never clarified – gap between mere subjective probability and trust.

In Chapters 3 and 8 we will criticize the reduction of trust to subjective probability, arguing about the importance of explicitly distinguishing internal and external components of the prediction; and also explaining how crucial and inherent for real trust are the bases (beliefs) supporting the prediction. We do agree with Williamson (see Chapter 8): if trust is just a euphemism for 'subjective probability' we do not need this term, we already have a strong theory of it, and a new vague and merely evocative term is just a confusing layer. On the contrary, we believe that trust is a specific, well-defined, *mental and social construct*.

On the basis of this 'internal attribution' ('design' or 'intentional' stance) *foundation* of trust we are able to account for several things.

For example in Section 2.7.2 we argue that there is internal versus external trust and explain why it is important to differentiate them. Consider a user working (collaborating, negotiating, etc.) on the web with other users, she has to distinguish her trust in these other potential partners from her trust in the actual context: the infrastructure with its internal constraints and rules. We also show that trustee and context have different dynamics; etc.

However, it is important to make clear that a given trust is internal or external only 'relative' to the assumed entity. While changing the target, again the distinction applies and again a form of *internality* is needed. If we consider for example our trust in the technical infrastructure (that relative to the partner *Y*, was an 'external' condition), we are now necessarily doing some

'causal internal attribution' relative to its (good) working. In general: *if I trust an entity E (of any kind) I'm always ascribing to E some internal properties, some virtues, on which the possible success depends*, and I depend on these 'virtues'.

This 'internal attribution' foundation of trust explains why it is trivial and false that the failure of *Y* necessarily produces a decrement of *X*'s trust in *Y*, and a success of *Y* should necessarily increase or cannot reduce *X*'s trust in *Y*. The effect of the failure or of the success again depends on its 'attribution': How much and for which aspect is it ascribed to *Y*? For which aspect is it ascribed to an external (from *Y*) circumstance? Only internal attribution to *Y* affects trust *in* *Y*, since trust holds upon this; while an external attribution to *C* (say the environment, the infrastructure, etc.) obviously affects the trust in *C* (of course, it is really important also to understand the relation and correlation between *Y* and *C*) (see Chapter 6 for more details).

Mistrust and diffidence (Chapter 4) are *negative* forms of trust. They too entail an internal causal attribution of the inadequate or bad behavior of *Y*. It is for some internal 'virtue' that *Y* is poor or harmful; there is not simply something lacking such that I do not trust *Y* (enough); but, positively, I attribute to *Y* some 'defect'; I think something bad of him.

Trust as an *External Goal* on *Y*

When *X* trusts *Y*, an *external goal* is put on *Y* (Castelfranchi, 2000c). Moreover, *Y* is assumed to respond to this impinging goal:

(i) either, with an *internalization* of it; that is by an internal goal, copied by the external one; by 'goal adoption' (or goal-adhesion) (Section 2.8); of course, this is possible only if *Y* is an intentional agent;
(ii) or with some internal mediator of the external function; some 'mechanism', some 'functioning' satisfying/performing that function.

2.2.8 Trust, Positive Evaluation and Positive Expectation

Trust is not Reducible to a Positive Evaluation

That *trust* is a *positive evaluation* is also confirmed by the fact that expressing trust towards *Y* is a way of *appreciating Y*. To express trust in *Y* is an indirect/implicit positive evaluation act (even a *compliment*) towards him; and this can be highly appreciated, and is one of the reasons for *reciprocation* (Chapter 8).

However, (as we will see) trust cannot be reduced to a positive evaluation. This is so for two main reasons. First of all, because there is much more than evaluations in trust mental attitude: there are also other kinds of beliefs, for example, expectations. Second, a positive valuation about *Y* is not *per se* trust in *Y*, or a trust attitude towards *Y*. It is only a possibility and a *potential* for trust. Only relatively to a possible dependence of *X* on *Y*, and to a *possible* delegation/counting of *X* on *Y*, that evaluative beliefs become a trust attitude.

Given *X*'s goal that *Y* bring it about that $g_X \subseteq p$, as a means for *X*'s achieving g_X, the beliefs about *Y*'s *qualities* for this acquire the color of trust.

Good evaluations, positive beliefs, good reputation, are just *potential* trust. Trust also implies a (potential or actual) decision of 'counting on' *Y*, of risking something while betting on it. It cannot just be a positive evaluation where *X* is not concerned and involved. *X* can evaluate that *Y* has such a good quality (*X* likes it), or has done or even will do such a good action, but this doesn't mean that *X* trusts *Y* (for that quality or action): *X has to consider the possibility to rely on this action and quality for achieving some goal of hers* in the future. Trust is about the future and about *potentially* exploiting something (that's why *X* 'likes' *Y*). (See Section 2.6.2).

To be more precise, *a positive evaluation of Y is not 'trust' per se*, even as a simple trust attitude or disposition, because it must be considered within a possible *frame*. It is a matter of the 'gestalt' nature of complex mental states. The side of a square is a linear segment; but: is a segment the side of a square? Not per se; only if considered, imagined, within that figure, as a component of a larger configuration that changes its meaning/role. Analogously: trust is based on and implies a positive evaluation, and when there is not yet a decision/intention it just consists in this, but *only if viewed in a perspective of the potential larger mental state*. Given *X's* positive beliefs (evaluation) about *Y* (as for something), if it is the case *X* might decide to rely on *Y*. In this sense those evaluations are a pre-disposition to trust, a trust attitude towards *Y*.

The same holds for the 'prediction' about *Y's* behavior. It is not yet trust. It is trust only as a possible 'positive expectation', that is in relation to a goal of *X* and in the perspective of a possible reliance on it (see below).

Decomposing a 'gestalt' is not reducing it to its components.

So the correct representation of the trust 'core' would be the insertion of the basic square (Figure 2.4) within the broad potential phenomenon.

Notice that this can just be the view of an 'observer': I see in *X* a trust attitude, predisposition, potential, towards *Y*. Actually in *X's* mind there is just a good evaluation of *Y*.

We can arrive at a true/full trust *disposition* of *X* towards *Y* if this thought has been formulated in *X's* mind. *X* not only has a positive evaluation of *Y*, but she has explicitly considered this as a potential, a base for a possible (non excluded) delegation to Y: *'If I want I could count on Y as for. . .', 'If it will be the case I might rely on Y, since . . . .'*. (For a clear distinction – on such a basis – between mere potential attitude and a real 'disposition' see later Section 2.3.2).

This is the psychological relationship between a mere positive evaluation of *Y* and a positive evaluation as a trust component or basic trust attitude.

Trust as Positive Expectation

On the basis of her positive beliefs about *Y*'s powers and willingness (or actualization) *X* formulates a prediction about *Y*'s behavior and outcomes. This is why a lot of scholars define trust in terms of an *expectation*. However, an expectation is not simply a prediction (or a strong prediction).

So trust is a positive expectation. Where a *positive expectation* is the combination of a *goal* and of a *belief* about the future (prediction). *X* in fact both believes that *Y* will do the action α and desires/wishes/plans so. And she both believes and wants that the goal g_X will be realized (thanks to *Y*).[21] Moreover, *X* is 'expecting', that is, waiting and checking

[21] The fact that when *X* trusts *Y, X* has a positive expectation, explains why there is an important relationship between trust and hope, since *hope* implies some positive expectation (although weaker and passive: it does not necessarily depend on *X*, *X* cannot do anything else to induce the desired behavior); and why trust can be 'disappointed'.

for something. She is concerned with her prediction; and has the *goal to know* whether the expected event will really happen (Castelfranchi, 2005) (Miceli and Castelfranchi, 2002); (Lorini and Castelfranchi, 2003).

We have to introduce, briefly, our theory of expectations as peculiar mental representation, because it predicts and explains a lot of the features and behavior of 'trust' (like its complement or counterpart, like its exposure to disappointment, like its structural link and ambivalence towards fear (for example towards the authorities), and so on).

2.3 Expectations: Their Nature and Cognitive Anatomy

'Expectations' are not just 'predictions'; they are not fully synonyms. And we do not want to use 'expectations' (like in the literature) just to mean 'predictions', that is, epistemic representations about the future. We consider, in particular, a 'forecast' as a mere belief about a future state of the world and we distinguish it from a simple 'hypothesis'. The difference is in terms of *degree of certainty*: a hypothesis may involve the belief that future *p* is possible while in a forecast the belief that future *p* is probable. A forecast implies that the chance threshold has been exceeded (domain of probability).

Putting aside the degree of confidence (we need a general term covering weak and strong predictions), for us 'expectations' has a more restricted meaning (and this is why a computer can produce weather 'predictions' or 'forecasts' but does not have 'expectations'). In 'expectations':

(i) the prediction is *relevant* for the predictor; he is *concerned*, *interested*, and that is why
(ii) he is 'expecting', that is the prediction is aimed at being verified; he is *waiting* in order to know whether the prediction is true or not.[22]

Expectation is a suspended state *after* the formulation of a prediction.[23] If there is an expectation then there is a prediction, but not the other way round.

2.3.1 Epistemic Goals and Activity

First of all, *X* has the goal of knowing whether the predicted event or state really happens (epistemic goal). She is 'waiting for' this; at least out of curiosity. This concept of 'waiting for' and of 'looking for' is necessarily related to the notion of expecting and expectation, but not to the notion of prediction.

[22] Notice that the first two meanings of 'to expect' in an English dictionary are the following ones:

- *to believe with confidence, or think it likely, that an event will happen in the future*
- *to wait for, or look forward to, something that you believe is going to happen or arrive*

While the definition of 'to forecast' is as follows:

- *to predict or work out something that is likely to happen, for example, the weather conditions for the days ahead*

(Encarta® World English Dictionary © 1999 Microsoft Corporation). Notice, the second component of 'expecting' meaning (absent in 'forecasting'): *wait for, or look forward to*. But also the idea that there is some 'confidence' in expectation: the agent *counts on* that.

[23] 'Prediction' is the result of the action of predicting; but 'expectation' is not the result of the action of expecting; it is that action or the outcome of a prediction relevant to goals, the basis of such an action.

Either X is actively monitoring what is happening and comparing the incoming information (for example perception) to the internal mental representation; or X is doing this cyclically and regularly; or X will in any case at the moment of the future event or state compare what happens with her prediction (epistemic actions). Because in any case she has the goal to know whether the world actually is as anticipated, and if the prediction was correct. Schematically:[24]

$$(\textbf{\textit{Expectation X p}}) => (Bel_X{}^{t'}\ (\textit{Will-be-True}^{t''}\ p)) \wedge (Goal_X{}^{period(t',t''')}\ (\textit{KnowWhether}_X\ (p\ \textit{OR Not}\ p)^{t''})) \textit{ where } t''' \geq t'' > t'.$$

X has the expectation p if X believes (at the time t') that p will be true (at the time t'') and has the goal (for the period t'-t''') to know if p is true. This really is 'expecting' and the true 'expectation'.

2.3.2 Content Goals

The epistemic/monitoring goal described above (to know if p will be true) is combined with *Goal that p*: the agent's need, desire, or 'intention that' the world should realize. This is really why and in which sense X is 'concerned' and not indifferent, and also why she is monitoring the world. She is an agent with interests, desires, needs, objectives on the world, not just a predictor. This is also why computers, that already make predictions, do not have expectations.

When the agent has a goal opposite to her prediction, she has a 'negative expectation'; when the agent has a goal equal to her prediction she has a 'positive expectation' (see Section 2.3.5).[25]

In sum, expectations (*Exp*) are axiological anticipatory mental representations, endowed with *Valence*: they are positive or negative or ambivalent or neutral; but in any case they are *evaluated against some concern, drive, motive, goal of the agent*. In *Exp* we have to distinguish two components:

- On the one hand, there is a mental anticipatory representation, the belief about a future state or event, the 'mental anticipation' of the fact, what we might also call the pre-vision (to for-see).

The format of this belief or pre-vision can be either propositional or imagery (or mental model of); this does not matter. Here, the function alone is pertinent.

- On the other hand, as we have just argued, there is a co-referent goal (wish, desire, intention, or any other motivational explicit representation).

[24] We will not use here a logical formalization; we will just use a self-explanatory and synthetic notation, useful for a schematic characterization of different combinations of beliefs and goals.

[25] To be true a goal equal to the prediction in expectation is always there, although frequently quite weak and secondary relative to the main concern. In fact, when X predicts p and monitors the world to know whether it is actually p, she also has the goal that p, just in order to not disconfirm her prediction, and to confirm she is a good predictor, to feel that the world is predictable and have a sense of 'control'. (See Section 7.1.2). We are referring to *predictability*, that is, the cognitive component of self-efficacy: the need to anticipate future events and the consequent need to find such an anticipation validated by facts. This need for prediction is functional in humans in order to avoid anxiety, disorientation and distress. (Cooper and Fazio, 1984:17) have experimentally proved that people act in order to find their forecasts (predictions) validated by facts and feel distressed by invalidation.

Given the resulting *amalgam* these representations of the future are charged of value, their intention or content has a 'valence': it is positive, or negative.[26] More precisely, expectations can be:

- ***positive*** *(goal conformable):* $[(Bel_X^t\ p^{t'}) \wedge (Goal_X^t\ p^{t'})] \vee [(Bel_X^t \neg p^{t'}) \wedge (Goal_X^t \neg p^{t'})]$
- ***negative*** *(goal opposite):):* $[(Bel_X^t\ p^{t'}) \wedge (Goal_X^t \neg p^{t'})] \vee [(Bel_X^t \neg p^{t'}) \wedge (Goal_X^t\ p^{t'})]$
- ***neutral****:* $[(Bel_X^t\ p^{t'}) \wedge \neg (Goal_X^t\ p^{t'}) \wedge \neg (Goal_X^t \neg p^{t'})] \vee [(Bel_X^t \neg p^{t'}) \wedge \neg (Goal_X^t\ p^{t'}) \wedge \neg (Goal_X^t \neg p^{t'})]$
- ***ambivalent****:* $[(Bel_X^t\ p^{t'}) \wedge (Goal_X^t\ p^{t'}) \wedge (Goal_X^t \neg p^{t'})] \vee [(Bel_X^t \neg p^{t'}) \wedge (Goal_X^t\ p^{t'}) \wedge (Goal_X^t \neg p^{t'})]$ where $t' > t$.

2.3.3 The Quantitative Aspects of Mental Attitudes

Decomposing in terms of beliefs and goals is not enough. We need 'quantitative' parameters. Frustration and pain have an *intensity*, can be more or less severe; the same holds for surprise, disappointment, relief, hope, joy, ... Since they are clearly related with what the agent believes, expects, likes, pursues, can we account for those dimensions on the basis of our (de)composition of those mental states, and of the basic epistemic and motivational representations? We claim so.

Given the two basic ingredients of any Exp (defined as different from simple forecast or prediction) Beliefs + Goals, we postulate that:

> **P1:** *Beliefs* and *Goals* have specific quantitative dimensions; which are basically independent from each other.

Beliefs have strength, a degree of subjective certainty; the subject is more or less sure and committed about their content. *Goals have a value, a subjective importance for the agent.*

To simplify, we may have very important goals combined with uncertain predictions; pretty sure forecasts for not very relevant objectives; etc. Thus, we should explicitly represent these

[26] • Either, the expectation entails a cognitive evaluation. In fact, since the realization of p coincides with a goal, it is "good"; while if the belief is the opposite of the goal, it implies a belief that the outcome of the world will be 'bad'.

• Or the expectation produces an implicit, intuitive appraisal, simply by activating associated affective responses or somatic markers; or both.

• Or the expected result will produce a *reward* for the agent, and – although not strictly driving its behavior, it is positive for it since it will satisfy a drive and reinforce the behavior.

We analyze here only the expectations in a strong sense, with an explicit goal; but we mentioned expectations in those forms of reactive, rule-based behaviors, first in order to stress how the notion of expectation always involves the idea of a *valence* and of the agent being concerned and monitoring the world; second, to give an idea of more elementary and forerunner forms of this construct. It is in fact the case of proto-expectations or expectations in 'Anticipatory-Classifiers' based behaviors, strictly conceived as reactive (not really goal-driven) behaviors, but based on anticipatory representation of the outcomes (Butz and Hoffman, 2002), (Castelfranchi, Tummolini and Pezzulo, 2005), (Butz, 2002), (Drescher, 1991), (Pezzulo et al., 2008).

dimensions of goals and beliefs:

$$^{\%}Bel_X\ p^t; \quad ^{\%}Goal_X\ p^t$$

Where % in goals represents their subjective importance or value; while in beliefs % represents their subjective credibility, their certainty.

An *Exp* (putting aside the epistemic goal) will be like this:

$$^{\%}Bel_X\ p^t \wedge ^{\%}\ Goal_X\ [\neg]\ p^t$$

The subjective *quality* of those 'configurations' or macro-attitudes will be very different precisely depending on those parameters. Also, the effects of the invalidation of an *Exp* are very different depending on:

(i) the positive or negative character of the *Exp*;
(ii) the strengths of the components.

We also postulate that:

> **P2:** The dynamics and the degree of the emergent configuration, of the macro-attitude are strictly a function of the dynamics and strength of its micro-components.

For example, anxiety will probably be greater when the goal is very important and the uncertainty high, than when the goal is not so crucial or the certainty is high. Let us characterize a bit some of these emergent macro-attitudes.

Hope and fear. 'Hope' is in our account (Miceli and Castelfranchi, 2010), (Miceli and Castelfranchi, 2002) a peculiar kind of 'positive Exp' where the goal is rather relevant for the subject while the prediction is not sure at all but rather weak and uncertain.[27]

$$^{low}Bel_X\ p^t \wedge ^{high}\ Goal_X\ p^t$$

Correspondingly one might characterize being afraid, 'fear', as an *Exp* of something bad, i.e. against our wishes:

$$^{\%}Bel_X\ p^t \wedge ^{\%}\ Goal_X \neg\ p^t$$

but it seems that there can be 'fear' at any degree of certainty and of importance.[28]

Of course, these representations are seriously incomplete. We are ignoring their 'affective' and 'felt' component, which is definitely crucial. We are just providing their cognitive skeleton.

[27] To be more precise, 'hope' contains just the belief that the event is 'possible', not that it is 'probable'.

[28] To characterize *fear* another component would be very relevant: the goal of avoiding the foreseen danger; that is, the goal of *doing* something such that Not p. This is a goal activated while feeling fear; fear 'conative' and 'impulsive' aspect. But it is also a component of a complete fear mental state, not just a follower or a consequence of fear. This goal can be quite a specified action (motor reaction) (a cry; the impulse to escape; etc.); or a generic goal 'doing something' ("my God!! What can I do?!") (Miceli and Castelfranchi, 2005). The more intense the felt fear, the more important the activate goal of avoidance.

2.3.4 The Implicit Counterpart of Expectations

Since we introduce a quantification of the degree of subjective certainty and reliability of belief about the future (the forecast) we get a hidden, strange but nice consequence. There are other implicit opposite beliefs and thus implicit Exp. For 'implicit' belief we mean here a belief that is not 'written', is not contained in any 'data base' (short term, working, or long term memory) but is only potentially known by the subject since it can be simply derived from actual beliefs (see Section 8.2.1 for more details). See also Figure 8.3 and the following discussion for a more specific analysis about implicit expectations.

2.3.5 Emotional Response to Expectation is Specific: the Strength of Disappointment

As we said, the effects of the *invalidation* of an expectation are also very different depending on: a) the positive or negative character of the expectation; b) the strengths of the components. Given the fact that *X* has previous expectations, how does this change her evaluation of and reaction to a given event?

Invalidated Expectations

We call invalidated expectation an expectation that happens to be wrong: i.e. while expecting that *p* at time t', *X* now beliefs that *NOT p* at time t'.

$$(Bel_X{}^{t}\ p^{t'}) \wedge (Bel_X{}^{t''} \neg p^{t'})\ \ where\,(t'' > t)$$

This crucial belief is the *'invalidating'* belief.

- Relative to the goal component it represents 'frustration', 'goal-failure' (is the *frustrating* belief): I desire, wish, want that *p* but I know that *not p*.
  FRUSTRATION: $(\text{Goal}_X\ p^{t'}) \wedge (\text{Bel}_X \neg p^{t'})$
- Relative to the prediction belief, it represents 'falsification', 'prediction-failure':
  INVALIDATION: $(Bel_X{}^{t}\ p^{t'}) \wedge (Bel_X{}^{t''} \neg p^{t'})$; *where* $(t' > t)$ *and* $(t'' > t)$
  $(Bel_X{}^{t}\ p^{t'})$ represents the former illusion or delusion (*X* illusorily believed at time *t* that at t' *p* would be true).

This configuration provides also the cognitive basis and the components of '**surprise**': *the more certain the prediction the more intense the surprise.* Given positive and negative expectations and the answer of the world, that is the *frustrating* or *gratifying* belief, we have: the configuration shown in Table 2.1.

Disappointment

Relative to the whole mental state of 'positively expecting' that *p*, the *invalidating* & *frustrating* belief produces 'disappointment' that is based on this basic configuration (plus the affective and cognitive reaction to it):

$$\text{DISAPPOINTMENT} : (^{\%}Goal_X{}^{period(t,t')}\ p^{t'}) \wedge (^{\%}Bel_X{}^{t}\, p^{t'}) \wedge (^{\%}Bel_X{}^{t'} \neg p^{t'})$$

Table 2.1 Relationships between Expectation and Surprise

	p	***¬ p***
$(Bel_X^t\ p^{t'})^{t<t'} \wedge (Goal_X\ p^{t'})$	No surprise + achievement	*surprise* + *frustration* ***disappointment***
$(Bel_X^t \neg p^{t'})^{t<t'} \wedge (Goal_X\ p^{t'})$	*surprise* + *non-frustration* ***relief***	no surprise + frustration

At *t X* believes that at t' (later) *p* will be true; but now – at t' – she knows that *Not p*, while she continues to want that *p*. Disappointment contains goal-frustration and forecast failure, surprise. It entails a greater *sufferance* than simple frustration for several reasons: (i) for the additional failure; (ii) for the fact that this impacts also on the self-esteem as epistemic agent (Badura's 'predictability' and related 'controllability') and is disorienting; (iii) for the fact that losses of a pre-existing fortune are worse than missed gains (see below), and a long expected and surely desired situation are so familiar and 'sure' that we feel a sense of loss.

The stronger and well-grounded the belief, the more disorienting and restructuring is the *surprise* (and the stronger the consequences on our sense of predictability) (Lorini *et al.*, 2007). The more important the goal is, the more *frustrated* the subject.

In disappointment, these effects are combined: *the more sure the subject is about the outcome and the more important the outcome is for her, the more disappointed the subject will be.*

- The degree of disappointment seems to be a function of both dimensions and components[29]. It seems to be felt as a unitary effect:

'How much are you disappointed?' 'I'm very disappointed: I was sure to succeed'
'How much are you disappointed?' 'I'm very disappointed: it was very important for me'
'How much are you disappointed?' 'Not at all: it was not important for me'
'How much are you disappointed?' 'Not at all: I have just tried; I was expecting a failure'.

Obviously, worst disappointments are those which place great value on the goal and a high degree of certainty. However, the *surprise* component and the *frustration* component remain perceivable and a function of their specific variables.

Relief

Relief is based on a 'negative' expectation that results in being wrong. The prediction is invalidated but the goal is realized. There is no frustration but surprise. In a sense relief is the opposite of disappointment: the subject was 'down' while expecting something bad, and now feels much better because this expectation is invalidated.

$$\text{RELIEF} : (Goal_X \neg p^{t'}) \wedge (Bel_X\ p^{t'}) \wedge (Bel_X{}^{t'} \neg p^{t'})$$

[29] As a first approximation of the degree of disappointment one might assume some sort of multiplication of the two factors: Goal-value * Belief-certainty. Similarly to 'Subjective Expected Utility': the greater the SEU the more intense the Disappointment. * = multiplication.

- *The harder the expected harm and the more sure the expectation (i.e. the more serious the subjective threat) the more intense the 'relief'.*

More precisely: the higher the worry, the threat, and the stronger the relief. The worry is already a function of the value of the harm and its certainty.

Analogously, **joy** seems to be more intense depending on the value of the goal, but also on how *unexpected* it is.

A more systematic analysis should distinguish between different kinds of surprise (based on different monitoring activities and on explicit versus implicit beliefs), and different kinds of disappointment and relief due to the distinction between 'maintenance' situations and 'change/achievement' situations.

More precisely (making the value of the goal constant) the case of loss is usually worse than simple non-achievement. This is coherent with the theory of psychic suffering (Miceli and Castelfranchi, 1997) that claims that pain is greater when there is not only frustration but disappointment (that is a previous *Exp*), and when there is 'loss', not just 'missed gains', that is when the frustrated goal is a maintenance goal not an achievement goal. However, the presence of *Exp* makes this even more complicated.

2.3.6 Trust is not Reducible to a Positive Expectation

Is trust reducible to a positive expectation? For example, to the estimated subjective probability of a favorable event? (as in many celebrated definitions). Trust as belief-structure is not just an 'expectation' (positive/favorable).

Let us put aside the fact that trust (at least implicitly) is trust *in* an agent; it is an expectation grounded on an 'internal attribution'. Even not considering 'trust that Y will ....' or 'trust *in* Y', but just 'trust that p' (for example, 'I trust that tomorrow it will be sunny') there is something more than a simple positive expectation. X is not only positively predicting, but is 'counting on', that p is *actively* concerned; *X has something to do or to achieve, such that p is a useful condition for that*. Moreover, such an expectation is rather sure: the perceived favorable chances are greater than the adverse ones, or the uncertainty (the plausible cases) is assumed as favorable. This is one of the differences between *trust* and *hope*; the difference between 'I trust that tomorrow will be sunny' and 'I hope that tomorrow will be sunny'. In the second one, I'm less certain, and just 'would like so'; in the first one, I am more sure about this, and that is why I (am ready to) *count on* this.

In fact, even non-social trust cannot be simply reduced to a favorable prediction. This is even clearer for the strict notion of 'social trust' (*'genuine' trust*) (Section 2.11): which is based on *the expectation of adoption*, not just on the prediction of a favorable behavior of Y.

2.4 'No Danger': Negative or Passive or Defensive Trust

As we said, in addition to *Competence* and *Willingness*, there is a third dimension in evaluating the trustworthiness of Y: Y should be perceived as not threatening, as *harmless*.

Either Y is *benevolent* towards X (for similarity, co-interest, sympathy, friendship, etc.), or there are strong internal (moral) or external (vigilance, sanctions) reasons for not harming X. This very important dimension appears to be missing in the definitions considered by

(Castaldo, 2002); however, it is present in others, for example those from social psychology or philosophy.[30]

Perhaps the most 'primitive' and original component (nucleus) of trust (especially of implicit and affective trust) is precisely the belief or feeling: 'no harm here/from . . .' and thus to *feel safe*, no *alarm*, no *hostility*, being *'open to'*.., *well disposed*. This is why trust usually implies no suspect, no arousal and alarm; being accessible and non-defended or diffident; and thus being relaxed. The idea of '*no danger*' is equivalent to '*the goals of mine will not be frustrated by Y'*; which – applied to animated entities (animals, humans, groups, and anthropomorphic entities) – is specified as the idea that '*Y has no the goal of harming me*'.

We call this elementary form of trust: α*-form* (Negative or Passive or Defensive Trust). In a sense 'feeling safe' can be the basic nucleus of trust and entire in itself; seemingly without any additional component. However, looking more carefully we can identify the other core components. Clearly positive *evaluations* and *expectations* (beliefs) are there. If I don't worry and do not suspect any harm from you, this means that I evaluate you positively (good for me; not to be avoided; at least harmless), since not being harmed is a goal of mine. Moreover, this feeling/belief is an expectation about you: I do not expect damage from you; which is a passive, weak form of positive expectation. Perhaps I do not expect that you might actively realize an achievement goal of mine; but I at least expect that you do not compromise a maintenance goal of mine: to continue to have what I have.

It is rather strange that this basic meaning of 'trusting' and this component is not captured in those definitions (except indirectly, for example, with the term 'confidence',[31] or marginally). This is for us the most 'primitive' and basic nucleus of trust, even before relying on *Y* for the active realization of a goal of *X*; just *passively* relying on *Y* to not be hostile or dangerous, non harming *X*.

Of course there is a stronger, richer, and more complete form of trust (β*-form*: that we call 'active', 'positive', 'achievement' trust) not only due to the idea/feeling (expectation) that the other will not harm me (lack of negative expectations); but including specific positive expectations: the idea/feeling that the other *will 'adopt' (some of) my achievement goals*, will be helpful; that *Y*'s attitude is 'benevolent' in the sense that not only is it not hostile, noxious or indifferent, but that he can be disposed to adopt and realize my goal (or at least that *Y* can be useful for achieving my goals). I can count on *Y*, and make myself dependent on *Y* for realizing (some of my) goals.

In terms of the theory of 'interference' (the basic notion founding social relations and action (Castelfranchi, 1998), α*-form* is the assumption or feeling that 'there can/will not be *negative interferences*' from/by *Y*'s side; while β*-form* is the assumption or the feeling that 'there can/will be *positive interferences* from/by *Y*′ (Where 'by' means 'on purpose': goal-oriented or intentional).

To be 'full' and complete trust should contain both ideas (α*-form*) and (β*-form*); but this is not always true. Sometimes it is more elementary and seems just limited to (α*-form*); sometimes it is mainly (β*-form*).

[30] See, for example, (Hart, 1988): trust enables us to assume "benign or at least non-hostile intentions on the part of partners in social interactions".

[31] The English term *'confident'/'confidence'* seems mainly to capture this nucleus of trust.

In a sense – as we have said – *α-form* is always present, at least implicitly; also when there is the richer and more specific attitude of *β-form*. In fact, when applied to the same set of goals (*β-form*) implies (*α-form*):

> *if Y will be useful or even helpful for the achievement of the goal g_X, he is not a problem, a threat for the goal g_X.*

To favor goal g_X implies to not harm goal g_X, since the achievement of goal g_X implies the non frustration of goal g_X. However, this implication does not mean that when X trusts Y (as capable and responsible for certain goals) X is always and fully relaxed and has nothing to worry from Y. In fact: what about other possible goals of X?

Except when applied to the same sub-set of goals or when 'generalized' (i.e., applied to all X's possible goals) *β-form* in fact doesn't necessary imply *α-form*. If trust is relative to a sub-set of X's goals, it is perfectly possible that X trusts Y (in the sense of *β-form*) for certain goals, but X could worry about her other goals; or, vice versa, that X trusts Y (in the sense of *α-form*) which is unwarlike, not threatening, but X cannot trust (in the sense of *β-form*) him as able and helpful towards some goal of hers. Thus, *β-form* and *α-form* don't necessarily co-occur, except for the same subset of goals. To be true the *α-form* potentially entails the *β-form* since it is the presupposition for a possible reliance.

2.5 Weakening the Belief-Base: Implicit Beliefs, Acceptances, and Trust by-Default

To make things simpler, we assume in our model that trust is composed by and based on 'beliefs'. However, this is an *antonomasia*: trust is based on *doxastic attitudes*: beliefs, knowledge, but also just *acceptances* (in our vocabulary: *assumptions*). Beliefs are assumed to be true in the world; to match with the (current, future, previous) world, if/when tested; or at least they are produced with this function and perspective. But we also have different and weaker doxastic attitudes on mental representations; or better different functions and uses of them.

For example, a very important function for the theory of purposive behavior (and for the theory of trust) is the use of doxastic representations as *conditions* for actions and decisions. In order to decide to act and to act (and for choosing an action) some conditions must be true, or better: they must be *assumed* (but not necessarily verified or proved). These are *assumptions*. We can use beliefs as assumptions; but they can also be unreal beliefs.

We can base our actions or reasoning on simple or mere 'assumptions' (non-belief assumptions), which have not been tested or are not destined to be tested. They are just – implicitly and automatically or explicitly – 'given for granted' or used 'as if'. Only the success of the practical action based on them, will provide an unconscious and indirect feedback about their 'truth'; will 'confirm' and indirectly 'verify' them. *It is important to distinguish between mere-assumptions and beliefs because one cannot decide to believe something while one can decide to assume something* (Cohen, 1992) (Engel, 1998).

This is also very relevant for trust because sometimes we trust Y not on the basis of real – more or less certain – beliefs (based on experience or on inference), but just assuming something about Y, and acting 'as if'. It is even possible to explicitly 'decide' to trust Y. We do not have sufficient evidence; current evidence does not provide us with enough certainty ('degree of trust') for trusting Y, but we can overcome this situation not by waiting or searching for

additional information and evidence, but just by making our decision: 'I have decided to trust you, (although . . .)'. This does not only mean that I have decided to rely on you (trust as *act*), but it can also mean that I am assuming something about you (expectations, evaluations), and that I am testing you out: precisely those assumptions will be confirmed or invalidated by your behavior (see Section 2.6). Sometimes we even delegate a task to *Y*, we take a risk, in order to get information about *Y*'s trustworthiness; we put him on test.

Trust beliefs obviously can be just *implicit*; not explicitly represented and considered in *X*'s mind. They can be just presupposed as logical conditions or logical entailment of some (explicit) belief. Suppose, for example, that *X* has such an evaluation about *Y*: '*Y* is a very good medical doctor' and suppose that this evaluation comes from *Z*'s recommendation (and *X* trusts very much *Z*), or from her direct practical experience. In this evaluation *X* implicitly assumes that *Y* is well prepared (competent on the subject), and also technically able to apply this doctrine, and also reliable in interaction, he takes care of you and of your problems. All these evaluations, or these possible pieces of *X*'s trust in *Y* (the fact that *X trusts Y for* being prepared, for taking care of, etc.; the fact that *X trusts in Y*'s expertise, attention, etc.), are just implicit beliefs; not necessarily explicitly derived and 'written' in *X*'s mind (just 'potential' (Levesque, 1984; Castelfranchi, 1996; 1997)), and/or not explicitly focused and taken into account.

It is also important to remark that there are many forms of trust which are not based on such an explicit and reason-based (argumentative) process we have presented in previous sections. They are rather automatic; not real 'decisions' or 'deliberations'. These are forms of *Trust* 'choices' and 'acts' based on some 'default rules' (positive evaluations are just implicit). The rule is: 'Except you have specific signals and specific reasons for do not trusting *Y* and rely on him, trust him'.

So the lack of distrust is the condition for trusting more than the explicit presence of trust evaluations (see Table 2.2; where 'Not (Believe q)' denotes the absence of such Belief in *X*'s mind).

These forms of automatic or by-default trust are very important, not only for characterizing generalized dispositions of people or affective trust attitudes, but also in other domains. For example, trust in our own natural information sources (our memory, our eyes, our reasoning) and, frequently even in social information sources, is suspect-less, is automatic, by-default, and doesn't need additional justification and meta-beliefs. (Castelfranchi, 1997).

Moreover, to 'believe' is in general not a real 'decision', but is certainly the result of some computation process based on sources, evidences, and data. But – in a sense – to 'come to believe' is an act of trust:

- trust in the belief (you rely on it and implicitly assume that it is valid, true), and
- Implicitly trust in the sources and bases of the belief; and
- (just procedurally and fully implicitly) even trust in the machinery or 'procedure' which outputs the belief.

Table 2.2 By-default Trust

IF	*Not (Believe X Not (Trustworthy Y τ C))*
THEN	*Trust (X Y C τ g_X) will be 'naturally' over the threshold for delegating*

2.6 From Disposition to Action

As we said, trust cannot be limited to a (positive) evaluation, an esteem of *Y*, and to a *potential* disposition to relying on him. *This potential can become an act*. On the basis of such a valuation and expectation, *X* can decide to entrust *Y* within a given 'task', that is to achieve a given goal thanks to *Y*'s competent action. 'To trust' is also a *decision* and an *action*.

The decision to trust is *the decision* to depend on another person to achieve our own goals; the free *intention* to rely on the other, to *entrust* the other for our welfare. However, to pass from a mere potential evaluation to a reliance disposition, that is, to the beliefs supporting the decision and the act to rely upon *Y*, the kernel ingredients we just identified are not enough.

At least a third belief (a part from that on being safe) is necessary for this: a *Dependence Belief*.

In order to trust *Y* and delegate to him, *X* believes that either *X* needs him, *X* depends on him (*strong dependence*), or at least that it is better for *X* to rely than to not rely on *Y* (*weak dependence*).

In other words, when *X* trusts someone:

- *X* has an active goal (not just a potential one; see Section 2.6.2); and
- *X* is personally and not neutrally 'evaluating' *Y*; moreover;
- *X* is in a strategic situation (Deutsch, 1985): *X* believes that there is 'interference' (Castelfranchi, 1998) and that her rewards, the results of her projects, depend on the actions of another agent *Y*.

To express it more clearly, we could say that:

> *Strong dependence* (Sichman *et al.*, 1994), is when *X* is not able at all to achieve her goal; she lacks skills or (internal or external) resources, while *Y* is able and in condition to realize her goal. *Y*'s action is a necessity for *X*.
>
> *Weak dependence* (Jennings, 1993), is when *X* would be able to realize her goal; however, she prefers to delegate to *Y*, to depend on *Y*. This actually means that *X* is strongly dependent on *Y* and needs *Y* for a broader outcome, which includes her original goal plus some additional advantage (like less effort, higher quality, etc.). This is why she prefers and decides to delegate to *Y*. In other words, *X* is reformulating her goal (which includes the original one: G_0), and, relative to this new goal, she is strongly dependent on *Y*. Then she formulates the instrumental sub-goal (plan) about '*Y* doing the action α', and – of course –for this goal also she strictly depends on *Y*.

These dependence beliefs (plus the goal g_X) characterize *X*'s 'trusting *Y*' or *'trust in Y'*[32] in delegation. However, another crucial belief arises in *X*'s mental state – supported and implied by the previous ones – the *Fulfillment Belief*: *X* believes that g_X will be achieved and *p* will

[32] We are stressing now the internal attribution of trust and putting aside for the moment the external circumstances of the action (opportunities, obstacles, etc.). We will analyze this important distinction further in 2.10 about *social trust*, and in Chapter 3 on decision.

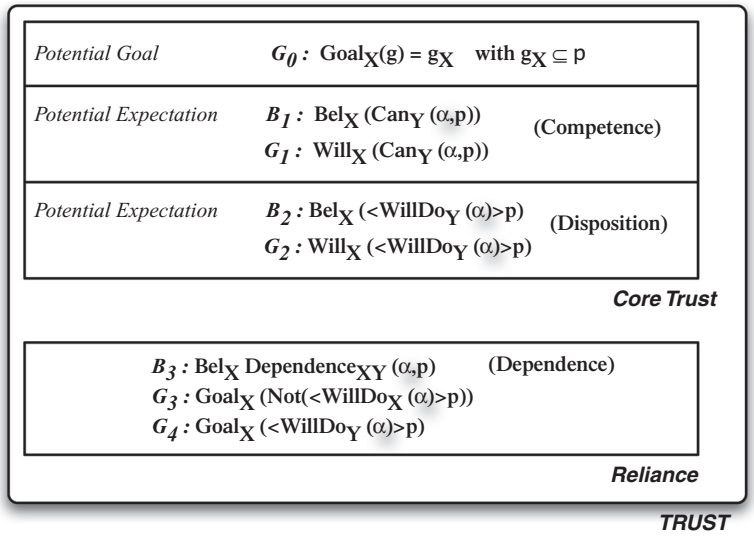

Figure 2.6 Mental State of the Decision to Trust

be true (thanks to Y in this case). This is the *'trust that' g_X (case in which $g_X = p$).* That is, X's trust in Y about task τ, for goal that p, implies also some trust that p will be realized (thanks to Y).[33]

When X decides to trust, X has also the new goal that Y performs α, and X relies on Y's α in her plan (delegation) (for more on this decision see Chapter 3). In other words, on the basis of those beliefs about Y, X 'leans against', 'counts on', 'depends upon', 'relies on'; *X practically* 'trusts' Y. Where, notice, 'to trust' does not only mean those basic beliefs (the core: see Figure 2.4) but also the decision (the broad mental state) and the act of delegating.

To be more explicit: *on the basis of those beliefs about Y, X decides of not renouncing to g_X, not personally bringing it about, not searching for alternatives to Y, and to pursue g_X through Y. This decision is the second crucial component of the mental state of trust: let us call this part reliance trust (we called the first part core trust) and the whole picture mental state of trust and the delegation behavior.*

Also, once more using Meyer, van Linder, van der Hoek *et al.*'s logics ((Meyer, 1992), (van Linder, 1996)), we can summarize and simplify the mental ingredients of trust as in Figure 2.6.

Of course, there is a coherence relation between these two aspects of trust (core and reliance): the decision of betting and wagering on Y is grounded on and justified by these beliefs. More than this: the degree or strength (see Chapter 3) of trust must be sufficient to decide to rely and bet on Y ((Marsh, 1994), (Snijders, 1996)). The trustful beliefs about Y (core) are the presuppositions of the act of trusting Y.

[33] Is Section 2.6.1 we will be more precise and complete on the relationships between "Trust that" and "Trust in".

2.6.1 Trust That and Trust in

As we have seen, both analyzing the *fulfillment belief* and the *attributional nature* of trust, there are two notions of trust which are orthogonal to the other fundamental distinction between trust as mere attitude (beliefs) and trust as decision and action. We refer to the distinction between *Trust that* and *Trust in*.

Trust that is our trust in a given desired state of the world, event, outcome; while *Trust in* is the trust in a given agent for doing something: in sentences like 'I trust John (for that/doing that)', 'I have/feel/put trust in John (for that/doing that)', 'I entrust John within that', etc.). Necessary and systematic relationships have emerged between these two forms of trust: they imply each other.

On the one side, *Trust in Y* (both as an evaluation and potential expectation and decision, and as an actual decision and action) necessarily entails the *Trust that* the action will be performed, and –thus – the trust that p will be true, the goal g_X will be realized. In other words, the 'Trust in Y' is just the trust that Y will 'bring it about that p'.

Even more analytically: the *trust in Y*'s qualities: competence, willingness, etc. entails the *trust in Y* as for doing α, which entails that α will be correctly performed, which entails that the goal will be realized (p will be true).

But, on the other hand, as we said, any *trust that* a given event will happen, as trust, is more than a mere, quite firm and positive expectation, not only because X counts on this (she is betting on this for achieving something) but because it is based on the idea of some (although vague) active *process* realizing that result.

Any *Trust that* presupposes some *Trust in* some natural or social *agent*. Any '*Trust* that p will be true' (not just hope, not simply an expectation) presupposes a trust that some Y will bring it about that p.

An interesting example is Lula (the president of Brazil) interviewed before the vote with other candidates. While other candidates were saying 'I wish' 'I hope' or even 'I'm sure to win', Lula's response was: 'I trust to win'. The difference with the other sentences is not only about the degree of certainty ('hope' and 'wish' are weaker or less specified) or about the positive expectation ('sure' has a high degree but might also be negative); the difference is that while saying 'I trust that the result will be this' Lula is implicitly saying: 'I trust people, the voters' 'I trust my party', 'I trust myself as for being able to convince and attract'.

If I conceptualize the relation as 'trust', I implicitly assume that the result depends on some entity or process (some *agency*) and the trust is 'in' such an entity, which (I trust that) will bring about that p. So we can introduce a new operator *TRUST-That*, with just two explicit arguments (X and p) and show how it can be translated in the former *TRUST* operator:

$$\textbf{\textit{TRUST-that}}\ (X\ p)\ \textbf{\textit{implies TRUST}}\ (X\ Y\ C\ \tau\ gx) \tag{2.2}$$

where X believes that there is any active Y (maybe he is unknown to X) that will bring it about that p through a behavior/action/task (maybe they are unknown to X too). In any case, X trusting the final achievement p, trusts also the agent performing the task.

We claim that this is true even with natural agents, in sentences like: 'I trust that it will rain'; there is something more than in 'I'm sure that . . .', 'I hope that . . ..'; there is an implicit trust in some vague causal process that will produce the result 'it will rain'. On the other hand, as

seen above:

$$\textbf{\textit{TRUST}}\,(X\ Y\ C\ \tau\ gx)\ \textbf{\textit{implies TRUST-that}}\,(X\ p) \tag{2.3}$$

The relation between 'Trust in' Y and in his action and trust that p is just one and the same relation (or better reflects the same basic action-theory relation) between the *Intention to do* a given action and the *Intention that* a given result holds: when I have the *Intention* that something holds, this necessarily implies that I believe that it depends on me, I have to act, and I *Intend to do* something; vice versa, if I have the *Intention to do* an action, this is certainly in order to achieve a given goal/result; thus I also *Intend that* this result be realized. When all there are not up to me, but delegated to another agent Y, we get all these forms of *Trust that* (outcome and performance) and the prerequisites of the trust *in* the agent, his virtues, and the virtues of his action. In sum, we can say that:

$$\textbf{\textit{TRUST}}\,(X\ Y\ C\ \tau\ gx) <==> \textbf{\textit{TRUST-that}}\,(X\ p) \tag{2.4}$$

where <==> means 'implies and is implied by'. There is a bidirectional relationship between *Trust-that* and *Trust-in*.

2.6.2 *Trust Pre-disposition and Disposition: From* Potential *to* Actual *Trust*

Frequently in Chapter 1 and in this chapter (and also later), for the sake of simplicity, we have been identifying and collapsing the notion of trust 'attitude' and 'disposition'. However, one might distinguish between the mere beliefs about Y (evaluations and expectations) and a true 'disposition', *which is something more than an evaluation*, but it is also something less than the actual decision and act. It is something in between; and preliminary to the actual decision.

Trust disposition[34] is the *potential* decision to *trust*, or better, the decision to possibly (en)trust Y. Not only X evaluates Y, but *she also perceives this evaluation as sufficient for* (if/when needed) trusting Y and relying on him. 'If it would be the case/when there will be the opportunity ... I will trust Y', 'One/I might trust Y'. X is *disposed* to trust Y (if/when it will be the case).

In *Trust disposition* in a strict sense the *expectations* are also conditional or *potential* (not actual); X has not the actual goal that Y does a given action and realizes a given result. X only has the prediction that 'if she would have the goal that p, Y will/would realize it; she might successfully rely on Y'. X is not actually concerned, actually waiting for something (expecting); she is only *potentially* expecting this, because she only potentially has the goal. Only this *potential* reliance actually makes the mere evaluation a possible expectation, a trust attitude (Section 2.2.1).

This specific mental attitude (*Trust disposition*) is in fact very important also for both selecting and ordering the preferences of the trustor; they contribute to a decision on which context and environment she has to situate herself in the (near or far) future. If Mary can decide to live in different environments where different goals of hers will be supported (she believes

[34] We will not consider the other notion of 'disposition' relevant for trust. The idea of a personality trait, or of mood, which make us generally open, well disposed, trustful towards the others; and increase the probability that we trust Y. This is the notion traditional in social psychology, and used – for example – in McKnight's model (McKnight and Chervany, 2001).

so) by different (artificial, human, institutional, and so on) trustees, she will decide to live in one in which her more relevant (with higher priority) potential goals will be better supported.

So we can say that *Trust disposition* is in fact a real, very important regulator of agents' behavior and goal selection.

Even the *decision* to trust can be conditional or hypothetical; I have already decided that, I have the future directed intention: '(if it will be the case) to address myself to Y, to trust, to rely on Y'. I have already decided, but, actually, I am not trusting him (as decision in action and act).

So, Figure 2.1, on trust stages and layers, should be even more articulated: in a potential trust attitude (*(pre)disposition*) versus a richer trust attitude (*disposition*) contained in a trust decision. We have to sophisticate a bit our analysis of the cognitive attitude of trust, by explaining that such a nucleus evolves in fact from its preliminary stage (before the decision) to its inclusion in the decision.

It is important to realize that the *disposition* kernel of the decision, intention, and action of (en)trusting includes or presupposes the first kind and nucleus of trust that we have just characterized (evaluation, prediction) but is broader, or better it is the actualization of it.

We pass from a *potential* evaluation/expectation to an *actual* one. There is a difference between the mere preliminary and potential judgment 'One can trust Y', 'Y is trustworthy', and the executive prediction that Y will actually (in that circumstance) do as expected and will realize the goal. X passes from the beliefs **2.5** and **2.6** (and the Belief that q) to the additional and derived belief (**2.7**):

$$Bel_X < Can_Y(\alpha) >_p \tag{2.5}$$

that means: X believes that Y is able and in condition to do α and the result of this action would be p true in the world (which is a *positive evaluation* of Y and of the context); and

$$Bel_X < (q \rightarrow Do_Y(\alpha)) >_p \tag{2.6}$$

that means: X believes that there is a condition q able to activate the performance α of Y; if it will be the case, Y will do α[35] (a *prediction* and just a *'potential' expectation*, since not necessarily X, while evaluating Y relative to the goal resulting from α, do currently have such a goal (Miceli and Castelfranchi, 2000).

$$Bel_X < Will\text{-}Do_Y(\alpha) >_p \tag{2.7}$$

that means: X believes that Y will do the action and will achieve p (which is in fact combined with the active goal that $Do_Y(\alpha)$ and thus is a real *expectation*), and also contains the expectation of α and of its desired outcome p (the goal X is relying on Y for).

We call the formulas (**2.5**) and (**2.6**) the *potential* evaluation and mental attitude towards Y: *trust pre-disposition*; and the mental attitude towards Y and (**2.7**) in the decision to rely on him: *trust disposition*.

[35] Where q is something like: "if X will need his help"; "if X will make the request", "if it will happen that . . ..", and so on.

2.6.3 The Decision and Act of Trust Implies the Decision to Rely on

Let us now come back to the relations between trust and reliance. Consider Holton's very nice example of the drama course (Holton, 1994) (pp. 63–64): *'If you have ever taken a drama course, you have probably played this game. You are blindfolded. You stand in the middle of a circle formed by the others. They turn you round till you lose your bearings. And then, with your arms by your sides and your legs straight, you let yourself fall. You let yourself fall because the others will catch you. Or at least that is what they told you they would do. You do not know that they will. You let yourself fall because you trust them to catch you'.*

We would like just to add to Holton's analysis a more subtle distinction. To decide to let yourself fall down is not the same as deciding to trust them. You can decide to let yourself fall down even if you do not trust them at all; you believe that they want to play a trick and to make fun of you, and you are ready to protect yourself at the last moment. If you decide to trust you *not only* decide to let yourself fall, but you *decide to count on them*, to act assuming that they will catch you. You decide to do your part of the plan *relying on* them doing their part of the plan. (Moreover – in 'genuine' social trust – you would count on them because you count on their motivations and their social-adoptive attitude towards you, and – following Holton – assuming a 'participant stance' towards them (p. 66) as persons treating you as a person).

Deciding to attempt, to try and see, is not deciding to rely/count on, and this is necessary in 'deciding to trust'; although also to decide to rely on is not enough for deciding to trust.

For many authors 'trust' is only social (and in a deep sense of 'social', Section 2.11); and they try to disentangle 'trust' from 'reliance' just on such a basis. See again (Holton, 1994) (p. 68): *'I have reason for simple reliance on an object if I need something done and reliance on it is my best bet for getting it done; likewise for simple reliance on a person. But in cases which involve not just simple reliance but trust, my reasons can be more complicated. Just because trust involves moving to a participant stance, I can have further reasons to trust, since that move can itself be something I value. Suppose we are rock climbing together. I have a choice between taking your hand, or taking the rope. I might think each equally reliable; but I can have a reason for taking your hand that I do not have for taking the rope. In taking your hand, I trust you; in so doing our relationship moves a little further forward.(..)'*[36]

In that statement Holton seems very close to (Baier, 1986), which claims that *trust must be distinguished from mere reliance, because it is a special kind of reliance: reliance on a person's goodwill towards me.* (p. 234)[37]

We agree that trust must be distinguished from mere reliance, but in our view, the real distinction is not directly based on 'sociality': intentional stance, or the richer 'participant stance', good will, or moral stuff. There is a preliminary distinction, before arriving at the special form of 'genuine' trust (Section 2.11).

---

[36] The text continues like this: "*. . .. This can itself be something I value. We need not imagine that you would be hurt if I chose the rope over your hand; you might be perfectly understanding of the needs of the neophyte climber. But our relationship would not progress.*" This is a nice issue: why does the act of trusting Y creates or improves a positive relationship with him? We examine this issue in Chapter 6.

[37] Baier's claim is based on the example of some safety and reliance assured by threats and intimidation on *Y*. If I count on *Y*'s fear of me or of my bodyguards, or on their protection, I do not really 'trust' *Y*. We disagree on this, because we claim that there are different kinds and layers of social trust; the one based on 'good will' or benevolence is only a sub-case of the one based on goal-adoption towards me for whatever reason (even avoiding sanctions or revenge) (See Section 2.8).

In natural language, I can 'trust' even the rope or the rock, but this is more than just 'relying on' it or deciding to grasp it.

Trust is (conscious and free, deliberated) reliance based on a judgment, on an evaluation of Y's virtues, and some explicit or entailed prediction/expectation: 'How could you *trust* that rock? It was clearly so friable!' 'No, I have tested it before; I evaluated it and I was convinced that it would have supported me!'

What is Reliance?

As showed in Section 1.3, in any (intentional, external) action α there is one part of the causal process triggered by the action and necessary for producing the result/goal of the action and defining it which *is beyond the direct executive control of the Agent* (Ag) of α. In performing α, Ag is making reliance on these processes and this is true in both cases:

- if Ag knows this, models this in his mind, and *expects* this;
- if Ag doesn't understand the process, is not aware of it, or at least doesn't explicitly represent it in his plan (although he might be able to do so).

As we said (Section 1.3), in the first case reliance becomes *delegation*. *'Delegation' would be the subjective and chosen reliance*. Counting upon: conceiving in $Ag's$ individual mind a multi-agent plan including (planning, expecting) the action of another autonomous agent. In the second case we have pure reliance.

In Delegation (at least) one part of the delegator's subjective plan for successfully accomplishing the intentional act α and achieving its goal, is 'allocated' to another agent either natural (like the sun when bronzing; or a coffee to feel awake) or social (like a waiter to bring food).

Let us clarify the concept: *Ag is making reliance upon Y/P* (where Y is another agent and P is a process) when: *there are actions (or inactions) in Ag's plan which are based on/rely upon Y/P, which depend on it for their efficacy* (in other words: that process P due to Agent Y creates some conditions for the performance or for the efficacy of those actions), *and Ag decides to perform those actions or directly performs them, Ag invests on Y/P (costs), Ag risks, Ag is relying on the fact that P will actually happen.*

P (due to Y) is a necessary process/condition for the achievement of $Ag's$ goal, but it is not sufficient: *Ag has to do (or abstain from doing) something, and thus Ag has to decide something: whether counting on Y/P or not, whether investing on it; Ag has to take her own decision of exploiting it or not.*

'Delegation' requires some trust, and trust as free decision and action is about delegation. This also means that *trust implies that X has not complete power and control over the agent/process Y, he is relying and counting upon*. Trust is a case of limited power, of 'dependence'.

When Y is an autonomous cognitive agent this *perceived* degree of freedom and autonomy consists in its 'choice': Y can *decide* to do or not to do the expected action. With this kind of agent (social trust) we in fact trust Y for deciding, being willing, to do – against possible conflicting goals at the very moment and in the circumstance of the performance – what Y 'has to' do (for us); we trust (in) Y's motivation, decision, and intention.

This feature of trust strictly derives from our founding trust on reliance on a non-directly controlled process and agent, on the perception of this 'non-complete control', and risk; on the distinction between trust 'in' *Y*, and global trust (internal attribution); on the idea of 'delegation': of deciding to count upon such a process/agent. If I do not decide to depend on this, I do not care about its non-controllability.

Reliance is a much broader phenomenon than trust. It even covers cases where the agent is unaware of the needed mediation. Let us consider the various cases and degrees before trust.

a) *X* does not understand or know whether to rely on a given agent or process. However, the positive result – due also to *Y*'s action – reinforces and reproduces *X*'s behavior and his reliance on *Y*.[38] We can call this '*confidence*'.
b) *X* is aware of the contribution of *Y*, but he *doesn't decide* to rely on *Y*; it is just so. It is, for example, when I just become aware of my confidence and reliance in a given support; I realize that it is only thanks to *Y* (that physical support? The obscure work of that guy?) that my activity was possible and effective.
c) *X decides* to rely on *Y* (not necessarily because he trusts *Y*; even without trusting *Y*; for example it is obliged to).
d) *X decides to count on Y*, but *Y* is *not an autonomous agent*; *Y* doesn't decide 'to do' what *X* needs (for example, I rely on the fact that – after this – she will be tired; or I decide to rely/bet on the fact that tomorrow it will be sunny).
e) *X decides to rely* on *Y* because she *trusts Y* (autonomous agent), but *X* does not rely on *Y*'s adoption of his goal (not 'genuine' trust).

The 'act' of trust is not reducible to reliance; 'to trust' (as act) implies 'counting on', which implies 'to rely on', but is more than this.

'Counting on' is not just relying, it is first of all an (originally) conscious reliance; the agent knows to rely on a given process and entity. Moreover, this reliance is not simply the discovery of a state of fact, of a static given situation; it is a decision and the result of a decision, or at least a course of events, something that *X* expects will happen while 'doing' something. *X* is doing something (or deciding to do something) and she expects that this process will bring a good/desired result thanks to the action of another entity, that will create some necessary condition for the successful realization of the act or of the goal.

Counting on means to have in mind a multi-agent plan, where the *action* (the contribution) of *Y* is enclosed; and where *X* has to do her share, at least deciding to counting on, or deciding to do nothing (which is an action) and delegating and waiting for the result, or at least expecting for.[39]

Trust (as act) of course is not just counting on, it is counting on based on a good evaluation (of *Y's* capacity and predictability) and on a good expectation. Moreover, to *count on* may be weaker – as a degree of certainty – than *to trust*, or less 'free'; trust in an autonomous decision based on an internal evaluation or feeling. One might 'count on' something even when pushed,

[38] While *walking* actually I'm implicitly and unconsciously relying on the floor; until I do not have some nasty surprise.

[39] In other words: *Counting on* it is not just to 'delegate', but is to do my share since and until I assume that *Y* will do his own share. Delegating is (deciding to) allocate/assign an action to *Y* in order – then – to count on this. They are two complementary moves and attitudes; two faces of the same complex relation.

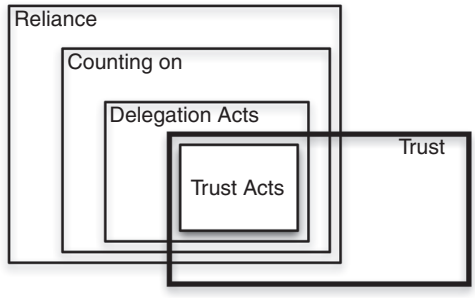

Figure 2.7 Relationships among Reliance, Counting on, Delegation, and Trust

obliged to do so, without really trusting it, and without a free decision.[40] If *X*'s choice was free she wouldn't count on *Y*, precisely because she does not trust *Y* (enough).

Delegation is the free act of counting on; and precisely for this reason it normally presupposes some trust. So, trust includes 'counting on' (but is not just reducible to it) which includes 'relying on' (see Figure 2.7).

2.7 Can we *Decide* to Trust?

In the next chapter we will model the decision process and how trust enters into it. However, in relation to our layered notion of trust, and to a belief-based account of trust, a crucial preliminary question arises. Can one *decide* to trust? For example, Annette Baier in her work on 'Trust and Antitrust' (Baier, 1986) claimed that we can never decide to trust (p. 235). We disagree on this.

First of all, the question is not well addressed without immediately distinguishing between trust as belief-structure (attitude/disposition) and trust as decision and action.

(i) As for *Trust-act*, the answer is 'yes': I can decide to trust somebody, and even when the background *belief-trust* wouldn't be sufficient. I can in fact say 'I have decided to trust him (to entrust him), although I do not really trust him'. I may in fact have other reasons and other values such that (even with a risk perceived to be too high) I *accept* that ('as if') *Y* be reliable and I decide to delegate and *relying on Y* and to be vulnerable by *Y*. For example, I might do so in order to see whether I am right or wrong, whether *Y* is really untrustworthy; or in order to show my good-faith and good-will; and I am disposed to pay such a price for having this proof or for exhibiting my virtues.[41]

(ii) As for *Trust as beliefs* about *Y* (evaluations and expectations), like for any other *belief* I cannot *decide* about.

[40] Suppose that you don't trust at all a drunk guy as a driver, but you are forced by his gun to let him drive your car.

[41] Consider that in this case (*trust-act* with not sufficient *belief-trust*) the trustor's mental elements will be conditioned by this act: for example, *X* will have, after the act and caused by it, the goal that *Y* is able and willing to achieve the task.

However, as we have already seen in Section 2.5, those assumptions might not necessarily and always be true/full beliefs, but just *acceptances*. I act on the presumption that, since I do not really know/assume that it is true; I do not know, but it might be. This is another meaning of the expression: 'I have decided to trust him': 'I have decided to give for credible, to give *credit*'; and on such a basis I have decided to rely on him.

The two decisions are not fully independent. On the one side, obviously, if I can *decide* to believe (assume) that you are trustworthy, then – on such a basis – I can decide to trust you, as an action of reliance. But there is also a strange relation in the other sense. Following Festinger's model of 'cognitive dissonance' reduction (Festinger, 1957), after a decision automatically and unconsciously we adjust the strength and the value of our beliefs on which the decision is based, in order to feel consonant and coherent (probably the function is to make our intentions more stable). So we increase the value of the beliefs favorable to our preference and choice, and make weaker the contrasting beliefs (focused on costs, risks, and the value of alternative choices). In terms of trust, *after the decision to trust Y (as intention and action) we will adjust the attitude/disposition beliefs (evaluations and expectations about Y).* So, the *decision* to *believe* obviously affects the *decision to trust* (counting on), but also the *decision to trust* may affect (feed back on) the *beliefs*.

Taking into account explicitly and consciously this effect in our decision-making would be irrational, since *Y*'s trustworthiness is not enhanced at all; only our subjective perception of it is enhanced. We can be (after the choice) less anxious, but not safer.[42] This is quite different from the other prediction about *Y*'s increased trustworthiness due to our action to trust him, as an actual effect on him (see Chapter 6).

2.8 Risk, Investment and Bet

Any act of trusting and counting on implies some bet and some risk. Trust is there precisely because the world is uncertain and risky (Luhmann, 1979). In fact, *X* is basing her behavior on uncertain expectations and predictions; and is making herself dependent on *Y*, and thus exposed to be vulnerable by *Y*. Also, because *Y* is not fully under *X*'s control; especially when he is an autonomous agent, with his own mind and interest.[43] *X* might eventually be disappointed, deceived and betrayed by *Y*: her beliefs may be wrong. At the same time *X* bets something on *Y*.

First, *X* renounced to (search for) possible alternatives (for example, other partners) and *X* might have lost her opportunity: thus *X* is risking on *Y* the utility of her goal g_X (and of her whole plan).

Second, *X* had some cost in evaluating *Y*, in waiting for its actions, etc. and *X* wasted her own time and resources.

Third, perhaps *X* had some cost to induce *Y* to do what *X* wants or to have him at her disposal (for example, *X* has paid for *Y* or for his service); now this investment is a real bet (Deutsch, 1985) on *Y*.

[42] We can take into account this expected effect in our decision, not as an increased reliability of *Y*, but as a good outcome of the decision, to be evaluated.

[43] Trust is "a device for coping with the freedom of others" (Gambetta, 1988) (p.219), or better with their "autonomy".

Thus, to be precise we can say that:

> *With the decision to trust Y (in the sense of relying on Y), X makes herself both more dependent and more vulnerable, and is more exposed to risks and harms by Y.*

2.8.1 'Risk' Definition and Ontology

What is risk? We accept the traditional view that risks for *X* are possible dangers impinging on *X* in a given context or scenario. More precisely the 'risk' is the estimated gravity/entity of the harm *multiplied for* its likelihood: the greater the possible harm the greater the risk; the greater the likelihood the greater the risk.[44]

However, we believe that this characterization is not enough and that it is important for the theory of trust (and for the theory of intentional action) to have a good **ontology of risks**, introducing several additional distinctions.

Any intentional action, any decision exposes, makes us *vulnerable*, to some 'risk'. Any act of trust, of relying on actions of others, exposes us to risks just for this general principle. The additional feature in trust is that the incurred risks *come from the others*: depends on their possible misbehavior; just because one has decided to depend on them. The decision to trust is the decision to not fully protect ourselves from possible dangers from *Y*, of exposing ourselves to possible dangers from *Y*, at least as for the disappointment of our positive expectation, the failure of the 'delegated' action, but possibly also the other risks resulting from our non diffident attitude, good faith, non protection. But other distinctions are needed.

First of all, it is crucial to distinguish between *'objective'* risks (the risks that *X* incurs following the point of view of an external ideal observer) from the *'subjective'*, perceived risks. Second, it is important to build on Luhman's intuition about the useful distinction between the dangers to which one is exposed independently from his decisions and actions, and those to which one is exposed as a consequence of his own decision and action. He proposes to call 'risks' only the second ones: the risks we 'take'.[45]

Given a certain point of choice in time, with different possible 'futures', different paths, there are risks which are only in one path and not on the others; risks which are in all possible future worlds[46]. When we choose to go in one direction or in another (by doing or not doing something) we 'take' those specific risks. But in a given path (or on all paths) there might be risks which do not depend at all on us and our choices. For example, if our planet would collapse under the tremendous impact of an asteroid (and this 'risk' is in fact there) this will

[44] Actually the two 'components' are not fully independent – from a psychological point of view. One dimension can affect the other: the perceived 'value' of the threatened goal or of the pursued goal can be modified by its chance (also, for example, for additional activated goals, like avoiding or searching for excitement, hazard); while the same estimated probability can be perceived/valued in different ways depending on the importance of goal.

[45] Although, as we will show, it is better to restrict the notion of "taken" risks, and of "taking risks" to a subset of this.

[46] This is not exactly the same distinction. There might be risks present on every path which are nevertheless dependent on the fact that we have 'chosen' to go in that direction; for example, the risk "to be responsible" for our action/choice and for some possible bad consequence.

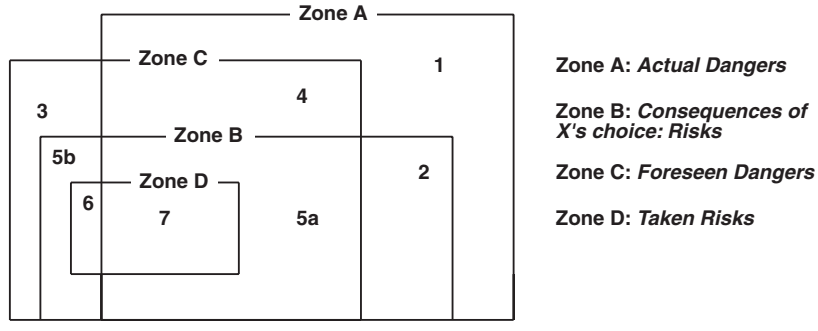

Figure 2.8 Risk Map

affect everybody living on earth no matter what their previous behavior was. It is important to combine these two dimensions: awareness and dependence.

The subject perceives, expects certain possible dangers: of course, he might be wrong. Thus some 'perceived', 'anticipated' risks are just 'imaginary': *X* believes that something is dangerous or that there is an incumbent risk, but he is simply wrong. Let's notice that frequently enough the lack of trust or distrust is due to an *over-estimation* of possible risks.

Some of those viewed, imagined risks are put aside, not taken into account in the decision, as unbelievable, too implausible, not to be considered. Others are, on the contrary, taken into account in and for the decision.

To 'take' a risk presupposes in our vocabulary that *X* assumes that such a risk is there and nevertheless decides to go in that direction. He knows (believes/assumes) and decides to expose himself to those dangers. Notice that *X* also 'takes' the imaginary risks; while he doesn't take the risks that don't realize at all, or that he has put aside. Trust is a subjective state and decision; thus what matters are not actual risks or safety, but the perceived safety and the believed risks. Not all the risks to whom *X* is exposed thanks to a given decision are 'taken'. We can resume the situation in the Figure 2.8.

This is where, about the specific intersections we have:

zone 1 represents *actual possible but unperceived dangers* <u>*not*</u> due to *X*'s choice

zone 2 represents *actual possible but unperceived dangers* due to *X*'s choice

zone 5 represents *Imagined (actual or unrealistic) dangers* not taken into account in the decision

zone 6 represents *Imaginary risks* (consequences of *X*'s choice) *evaluated in the decision, and thus "taken"*

zone 7 represents *Actual risks* (consequences of *X*'s choice) *evaluated in the decision, and thus "taken"*

zone 3 + **zone 5b** + **zone 6** represent *imaginary dangers*; perceived but not real

zone 4 + **zone 5a** + **zone 7** represent *perceived and realistic dangers*

Trust as decision has to do only with 'taken' risks: *perceived, imagined (true or false) risks to whom X believes to expose himself as a consequence of his decision/action.*

To 'feel confident' (we accept Luhman's proposal) has to do with any danger I (do not) feel exposed, independently from my own actions, in a given environment, context, situation (zones 3 and 4). However, trust as behavior and social phenomena (in economics, politics, etc.) also requires the theory of the 'objective' risks to which people or institutions are exposed thanks to their interdependence and reliance.

A particularly interesting case is risks X perceives (predicts) and in a sense 'chooses', but actually has no alternative: she has no real 'freedom of', she has no real responsibility in 'taking' that risk, since the alternative is even worst. Thus, even if X chooses a path which is not convenient at all, a bet which per se, in isolation, would be irrational, she is acting in a rational way: minimizing her risk and damage.

2.8.2 What Kinds of Taken Risks Characterize Trust Decisions?

When X trusts Y there are three risks:

a) the risk of failure, the frustration of g_X (*missed gains*) (possibly for ever, and possibly of the entire plan containing g_X);[47]
b) the risk of wasting efforts and investments (*losses*);
c) the risk of unexpected harms (frustration of others goals and/or interests of X).

As for the first risk (case *a*), the increment of X's dependence from Y is important.

Two typical cases are the dependence from time resources and trusted agents' resources. Maybe that after Y's failure there is no time for achieving g_X (it has a specific time expiration); maybe that initially X might have alternatives to Y (rely on Z or W) after her choice (and perhaps because of this choice) Z and W might be no more at her disposal (for example they might be busy); this means that X's alternative means (partners) for g_X are reduced and then X 's dependence on Y has increased (Sichman *et al.*, 1994).

Given those (in part perceived) risks and thus the explicit or implicit additional goals of avoiding these harms, X becomes – relative to these additional goals – *more 'dependent' on* Y, since actually it is up to Y (after $X's$ decision to trust him and relying on him) do not cause those harms to X.

As for becoming more vulnerable (case *c*), since X expects some help from Y (as for goal g_X) X feels well disposed towards Y. The (implicit) idea that there is no danger from Y (as for g_X), reduces X's diffidence and vigilance; X feels confident towards Y, and this *generalizes* beyond g_X. This makes X – less suspicious and careful – more accessible and undefended. This is also due to a bit of *transitivity* in Positive Trust from one goal to other: if X trusts Y for g_X, X can be a bit prone to trust Y as for a goal g'_X (where g'_X is different from g_X) even if we have to consider all the limits of the transitivity applied to the trust concept (see Chapter 6).

[47] Moreover there might be not only the frustration of g_X, the missed gain, but there might be additional damages as effect of failure, negative side effects: the risks in case of failure are not the simple counterpart of gains in the case of success.

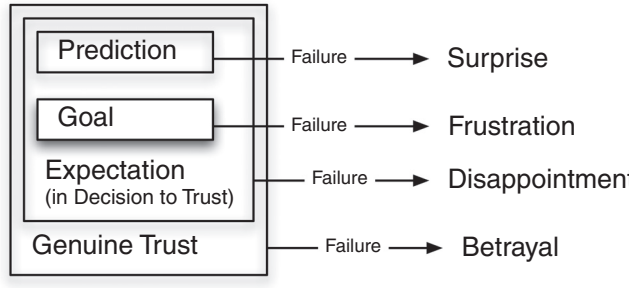

Figure 2.9 Potential Effects for the Failure of the Different Trust Elements

In sum, Y – after and because of X's trust (attitude + act) in him – can harm X in several ways:

- By frustrating X's goal (g_X) for which X relies on him. This also implies frustrating X's expectations and hope: that is *disappointing* X. And this will impact on X's self-esteem (as evaluator and/or decision maker).
- Y can also damage X's general attitude and feeling towards the world; and so on (Miceli and Castelfranchi, 1997).
- Moreover, X may, not only, be *surprised*, *frustrated*, *disappointed*, but X can feel *resentment* (and even indignation) for moral violations, for being *betrayed* (see Figure 2.9).
- By frustrating other goals of X that she did not protect and defend from Y's possible attack, being relaxed and non-diffident. This will imply analogous consequent frustrations.

It was necessary to immediately mention uncertainty and risk, in connection with the notions of reliance and expectations. However, we will deeply develop these issues (Chapter 3), after introducing the *degree* of certainty of beliefs, and the degree of trust in decision-making.

2.9 Trust and Delegation

What Delegation Is

As we said, *in Delegation the delegating agent (X) needs or likes an action of the delegated agent (Y) and includes it in her own plan: X relies, counts on Y. X plans to achieve g_X through Y. So, she is formulating in her mind not a single-agent but a multi-agent plan and Y has an allocated share in this plan: Y's delegated task is either a state-goal or an action-goal* (Castelfranchi, 1998) (see Figure 2.10).

To do this X has some trust both in Y's ability and in Y's predictability, and X should abstain from doing and from delegating to others the same task (Castelfranchi and Falcone, 1997).

We have classified delegation in three main categories: *weak*, *mild* and *strong delegation*.

(i) In *weak delegation* there is no influence from X to Y, no agreement: generally, Y is not aware of the fact that X is exploiting his action.

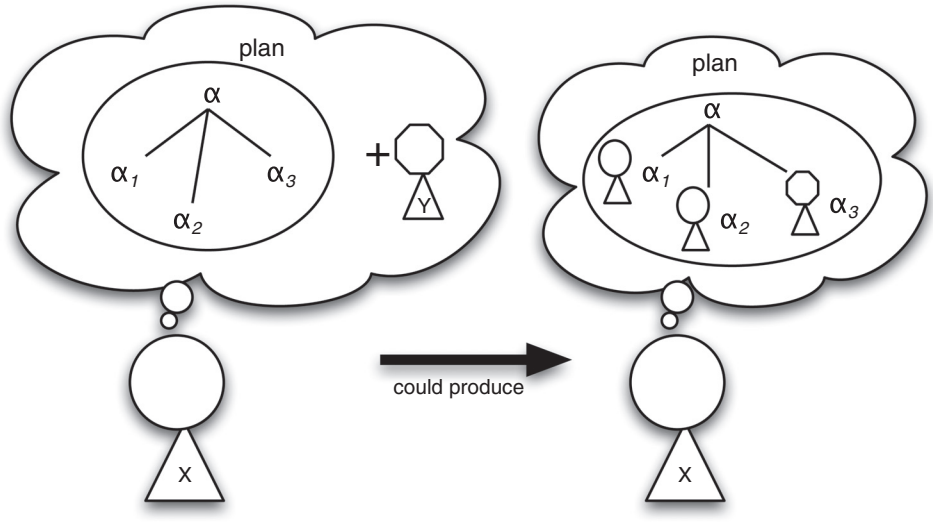

Figure 2.10 The potential for Delegation

As an example of weak and passive but already social delegation, which is the simplest form of social delegation, consider a hunter who is waiting and is ready to shoot an arrow at a bird flying towards its nest. In his plan the hunter includes an action of the bird: to fly in a specific direction; in fact, this is why he is not pointing at the bird but at where the bird will be in a second. He is delegating an action in his plan to the bird; and the bird is unconsciously and (of course) unintentionally collaborating with the hunter's plan.

(ii) In a slightly stronger form of delegation (*mild delegation*) *X* is herself eliciting, inducing the desired behavior of *Y* to exploit it. Depending on the reactive or deliberative character of *Y*, the induction is just based on some stimulus or is based on beliefs and complex types of influence.

(iii) *Strong delegation* is based on *Y*'s awareness of *X*'s intention to exploit his action; normally it is based on *Y*'s adopting *X*'s goal (for any reason: love, reciprocation, common interest, etc.), possibly after some negotiation (request, offer, etc.) concluded by some agreement and social commitment.

The Act of Delegation

Notice that weak delegation is just a mental operation or action, and a mental representation. *X*'s external action is just waiting for or abstaining from doing the delegated action or doing her own part of the plan. On the contrary, in (ii) and (iii) delegation is an external action of *X* on *Y*, which affects *Y* and induces him to do the allocated task. Here to delegate means to bring it about that *Y* brings it about that *p*. If $E_X(p)$ represents the operator '*to bring it about that*', indicating with *X* the subject of the action and *p* the state resulting of *X*'s action, we have the situation shown in Table 2.3.

Table 2.3 Role of X in the different Kind of Delegation

Weak Delegation	*X is exploiting* $E_Y(p)$
Mild Delegation	$E_X(E_Y(p))$
Strong Delegation	$E_X(E_Y(p))$

2.9.1 Trust in Different Forms of Delegation

Although we claim that trust is the mental counter-part of delegation, i.e. that it is a structured set of mental attitudes characterizing the mind of a delegating agent/trustor, however, there are important differences, and some independence, between trust and delegation. Trust and delegation is not the same thing.

Delegation necessarily is an *action* (at least mental) and the result of a decision, while trust can be just a potential, a mental attitude. The external, observable behavior of delegating either consists of the action of provoking the desired behavior, of convincing and negotiating, of charging and empowering, or just consists of the action of doing nothing (omission) waiting for and exploiting the behavior of the other. Indeed, we will use trust and reliance only to denote the mental state preparing and underlying delegation (*trust* will be both: the small nucleus and the whole).[48]

There may be trust without delegation: either the level of trust is not sufficient to delegate; or the level of trust would be sufficient but there are other reasons preventing delegation (for example prohibitions); or trust is just potential, a predisposition: 'X will, would, might rely on Y, if/when . . .', but it is not (yet) the case. So, trust is normally *necessary* for delegation, but it is not *sufficient*: delegation requires a richer decision.

There may be delegation (or better just 'counting on') without trust: these are exceptional cases in which either the delegating agent is not free (coercive delegation[49]*) or she has no information and no alternative to delegating, so that she must just make a trial (blind delegation).* So, all trust decisions and acts imply an act of delegation, but not every act of delegation is an act of trust.

Moreover, the decision to delegate has no degrees: either *X* delegates or *X* does not delegate. Indeed trust has degrees: *X* trusts *Y* more or less relatively to α. And there is a threshold under which trust is not enough for delegating.

Trust in Weak Delegation

While considering the possible temporal gap between the decision to trust and the delegation we have to consider some other interesting mental elements. (The temporal gap ranges between *0* and ∞; *0* means that we have delegation at the same time as the decision to trust; ∞ means that delegation remains just a potential action). In particular we have in all the cases (weak, mild and strong delegation) *X*'s *intention-that Y* will achieve the task (Grosz and Kraus, 1996). In every case this intention is composed through different intentions.

[48] In our previous works we used "reliance" as a synonym of "delegation", denoting the action of relying on; here we decide to use "reliance" for the (part of the) mental state, and only "delegation" for the *action* of relying and trusting.

[49] Consider the example of the drunk driver, in note 40.

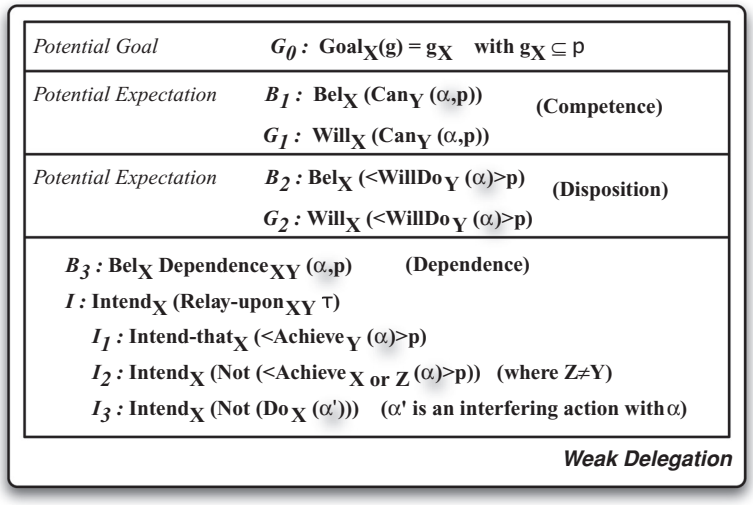

Figure 2.11 Mental Ingredients for Weak Delegation

In weak delegation, we have three additional intentions (I_1, I_2, and I_3 in Figure 2.11), respectively, the intention that Y achieves p by the action α; the intention to not to do (or do not delegate to others) that action; and the intention to not hinder Y's action with other interfering actions.

Trust in Mild Delegation

In mild delegation, in addition to I_1, I_2, and I_3 there is another intention (I_4), that is, X's intention to influence Y in order that Y will achieve τ (Figure 2.12).

Trust in Strong Delegation

In strong delegation X's intended action is an explicit request (followed by an acceptance) to Y about p (see Figure 2.13).

Consider that in mild and strong delegation the intentions are already present in the decisional phase and they are the result of an evaluation. For example, X has to evaluate if the delegation will be successful or not in the case of influence, request, etc.

2.9.2 Trust in Open Delegation Versus Trust in Closed Delegation

A very important distinction is also that between *open* and *closed* delegation. In *closed delegation* the task assigned to Y is fully specified. It is just a sequence of actions to be performed. The task is a merely executive task. The extreme case of this is the classical Tayloristic industrial organization of work, where the worker is explicitly forbidden to think

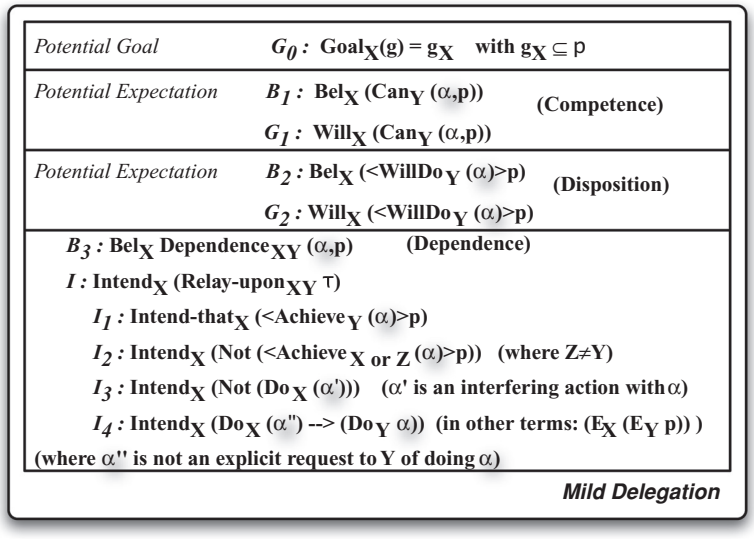

Figure 2.12 Mental Ingredients for Mild Delegation

about the delegated task (the conditions for realizing the task are constraining any potential free initiative in the job), and he has just to perform a repetitive and mechanic movement.

On the contrary, in (completely) *open delegation* the assigned task for Y is 'to bring it about that p', to achieve (in some way) a given result. Y has to find and choose 'how' to

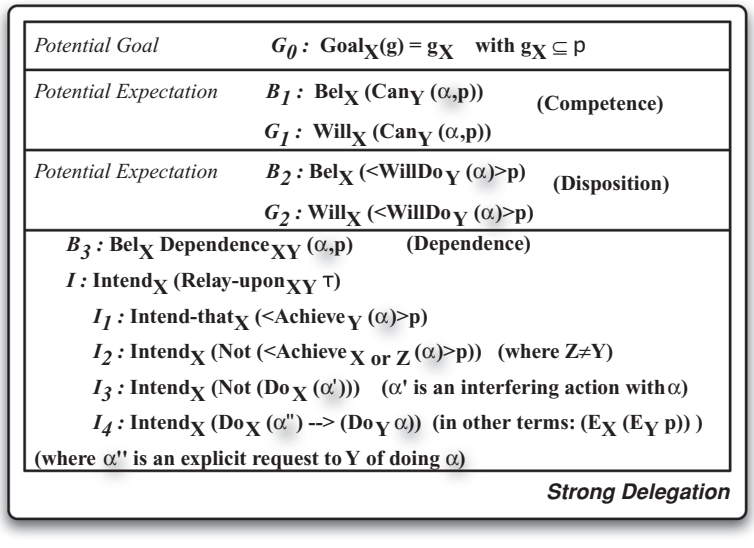

Figure 2.13 Mental Ingredients for Strong Delegation

realize his 'mission'. He has to use his local and timely information, his knowledge and experience, his reasoning, and so on. Of course, our evaluation and expectations about *Y* are very different in the two cases; and the trust that we have in *Y* refers to different things. In *closed delegation* we trust *Y* to be obedient, precise, and skilled; we trust him to 'execute' (with no personal contribution – if not minimal sensory-motor adaptation) the task, which is completely specified. In *open delegation* we trust *Y* not only in his practical skills, but also in his understanding of the real sense of the task, in his problem solving ability, in his competent (and even creative) solution.

We might even trust *Y* to violate our request and specific expectation in order to *over-help* us. Over-help (Castelfranchi and Falcone, 1998b) is when *Y* does more than requested (for example, he satisfies not only the required goal but also other goals of *X* in someway linked with it).

A special form of over-help is *critical help*: when *Y* is not able or in condition to do as requested, or understands that *X*'s request is simply wrong (for her goal), or he has a better solution for her problem, and violates her specific request but in order to satisfy her higher goal. This is the best form of real *collaboration* (even if risky in some way); *Y* is really helpful towards *X*'s goals; he is not a stupid executor.

Sometimes, we deeply trust *Y* for over-help or for critical-help. We confidently expect that – if needed – he will do more than requested or will even violate our request in order to realize the adopted our goal (or to guarantee our 'interests'). This is close to the most advanced trust: *tutorial trust* (Section 2.11.3).

2.10 The Other Parts of the Relation: the Delegated *Task* and the Context

2.10.1 Why Does X *Trust* Y*?*

It is not enough to stress that *X* trusts *Y* to do α (an action), or to do something. First of all, as we said, trust is not simply a (firm) prediction, but is a positive expectation. In other words *X* is interested in the performance of α, she expects some positive result; one of the outcomes of α is (considered as or is) a *goal* of *X*. As we have remarked, an aspect absolutely necessary, but frequently ignored, (or at least left implicit) is that of *the goal, the need,* relatively to whom and for the achievement of whom the trustor counts upon the trustee. This is implicit when some 'positive' result/outcome or the 'welfare' or 'interest' are mentioned in the definition, or some 'dependence' or 'reliance', or the 'vulnerability' are invoked. This is also why, when *X* trusts *Y* in something, he cannot only be surprised, but can be 'disappointed' (and even betrayed) by *Y* or by his performance. Moreover, as we have anticipated, the task allocated to *Y* – especially in *social trust* – is *delegated* to *Y*, since *Y* is an 'autonomous' agent, and in any case *X* is relying on a process which is beyond his power and control. Trust implies some (perceived) lack of controllability.

We call 'task' (τ) the delegated action α to be performed and the goal state *p* (corresponding or including g_X) to be realized by *Y* (in both cases in fact *Y* has to 'see it that . . .', to 'bring it about that . . .') (see Section 2.9.2 on *closed* versus *open* delegation), because it is something allocated to *Y* within a multi-agent plan; something *to be done*; something on which *X* counts

on; and something that frequently Y has an obligation to do (due to a promise, a role, etc.). The theory of *tasks* is important (see *trust generalization* in Chapter 6).

2.10.2 The Role of the Context/Environment in Trust

Trust is a context-dependent phenomenon and notion. This means that X trusts Y for τ on the basis of a specific context; just changing the context (for the same τ and the same Y) X's attitude and decision might be different.

Consider X's trust attitude towards the same agent Y for the same task τ when:

- he (Y) is in two completely different contexts (maybe with different environmental and/or social conditions);
- she (X) is in two completely different contexts (maybe with different environmental and/or social conditions).

In fact, one should perhaps be more subtle, and clearly distinguish these two kinds of context:

- the context of X's evaluation and decision (affecting her mind) while feeling trust for Y and deciding to trust him or not (*evaluation context*); and
- the context of Y's performance of α (*execution context*).

They are not one and the same context. The execution context affects Y's objective trustworthiness; his possibility to really achieve the goal in a good way; and – as perceived by X ($_XTW_Y$) – affects X's expectation.

But the evaluation context is the social and mental environment of X's decision. This can affect:

- X's mood and basic social disposition;
- X's information and sources;
- the beliefs activated and taken into account by X;
- X's risk perception and acceptance;
- X's evaluation of the execution context; and so on.

Moreover, the evaluation and decision of X also depends on this complex *environmental trust*: X's trust in the *environment* where α will be executed, which can be more or less interfering or harmful; in the *supporting infrastructure* (execution tools, coordination and communication tools, etc.); in the *institutional context* (authorities, norms, and so on); in the *generalized atmosphere* and *social values*; and so on.

Environmental trust (*external attribution*) and trust 'in' Y (*internal attribution*) must be combined for a decision; and they are also non-independent one from the other (see also Section 8.3.3 for evaluating the importance of this decomposition with respect the subjective probability).

Table 2.4 Conditional Trust

IF	*Event (e) = true*
THEN	*Trust (X Y C τ g_X) will be over the threshold for delegating*

Not only the trust in Y as for τ is context dependent, but if the context (environment) in the mental model of X plays an active causal role, X has also to trust the context, as favorable or not too adverse or even hostile (to Y or to τ). But, of course, Y's capacity and reliability may vary with the more or less adverse nature of the context: it might decrease or even increase. On this we have developed a specific section (see Chapter 6). This is also very important in trust dynamics, since it is not true that a failure of Y necessarily will decrease Y's perceived trustworthiness for X; it depends on the causal attribution of the failure. The same holds for success.

Another important way in which the context is relevant for trust, is that there can be different trusts about Y in different social contexts, related to the same task: for example, Y is a medical doctor, and he is very well reputed among the clients, but not at all among his colleagues. Or there can be different trusts in different social contexts because different *tasks* are relevant in those contexts. For example, Y is a well reputed guy within his university (as teacher and researcher), but has a very bad reputation in his apartment building (as an antisocial, not very polite or clean, noisy guy).

Trust can migrate from one task to another, from one trustor to another, from one trustee to another (see Chapter 6), and also from one social context to another. It depends on the connections between the two contexts: are they part of one another? Are they connected in a social network? Do they share people, values, tasks, etc.? So, trust is not only a context dependent and sensible phenomenon but is a context-dynamic phenomenon.

Moreover, not only is trust context-dependent but it can also be *conditional*: A special event (e) could be considered by X, in a given context and with respect to a specific trustee, as crucial for trusting Y (see Table 2.3).

Consider our example of a bus stop, in weak delegation. After Y raised his arm to stop the bus the driver is more sure that he will take the bus. In our view, this is not just simple 'conditional' probability (after *the first event*, or given condition C, the probability of *the second event* is greater or smaller). In real trust – given its attributional nature – the *first event* can be interpreted by X as a *signal*. For example, a given act or attitude or sentence of Y can be a sign for X of Y's capacity or of his internal disposition, which makes his doing τ more reliable.

2.11 *Genuine* Social Trust: Trust and *Adoption*

As we saw trust is not only a 'social' attitude. It can be directed towards an artifact or unanimated process. Someone would prefer another term, say *confidence*, but this is just a (reasonable) technical convention, not the real use and meaning of these words.[50] However, it is true that the most theoretically and practically relevant and the most typical notion of trust *is the social one*.

Social trust means *trust towards another autonomous agent perceived (conceived) as such.* That is, towards a purposive, self-governed system, not only with its own resources and

[50] Moreover, 'confidence' is very close to 'trust' in a non-technical meaning, it just seems to contain some reliance, and be quite social. It also seems to be based just on learning and experience.

causal-processes, but with its own internal control and *choices*. This is why social trust is there at all; or better, why social interaction requires trust.

The trustor cannot and does not fully 'control' (monitor and guide) *Y*'s activity towards the goal (*X*'s expected result). *X* passes to *Y* (part of) the *control* needed for realizing the expected result. *X* relies precisely on this. On the one side, this is precisely one of the main advantages of *delegating* (task assignment and counting on): *delegating also the control and governance of the activity* (even if *X* was able to perform it herself). But, on the other side, this is precisely *the specific risk of social trust*: not just possible external (environmental) interferences, but 'internal' interferences (due to *Y*'s nature and functioning).

Y is selecting the right action, employing resources, planning, persisting, executing; he might be defective on this. Moreover, since he has his own control system for purposive behavior, usually he has his own internal goals. Now, those *individual* goals may interfere; taking precedence, being in conflict and prevailing, etc.

If *X* decides to trust *Y*, to count on him, *X* expects (predicts and wishes) that *X*′*s* goal – adopted by *Y* – will prevail on *Y*'s autonomous goals, and will be pursued. This is the typical bet of social trust. *There is a peculiar relation between (social) trust and autonomy: we trust in autonomous systems and this is our specific (social) risk: possible indifference, or hostility, or changing of mind, or profiting and exploitation, up to a 'betrayal', which presupposes a specific or general, explicit or implicit, 'pact'*.

Since social trust is directed towards another *autonomous* agent, *considered as* an autonomous agent, with its attitudes, motivations (including the social ones), and some freedom of choice, it requires an *intentional stance* towards a social entity (with its own intentional stance towards us).

However, this is not yet enough to capture the most *typical* social notion of trust; what many authors (like Baier, Hardin, Holton, Tuomela) would like to call *genuine trust*. *Genuine* (social) trust, the basic, natural form of social trust, is based on *Y*'s *adoptive* attitude. That is, *X trusts Y's adoption of her interest/goal, and counts on this*. *Y* is perceived as taking into account *X*'s goals/interests; and possibly giving priority to them (in case of conflicts). This is true trust in a social agent 'as a social agent'.

Social goal-adoption, is the idea that another agent takes into account in his *mind* – in order to satisfy them– my goals (needs, desires, interests, projects, etc.); he 'adopts' them as goals of himself, since he is an 'autonomous agent', i.e. self-driven and self-motivated (but not necessarily 'selfish'!), and is not an hetero-directed agent, and can only act in view, be driven by, some internal purposive representation (Conte and Castelfranchi, 1995). So – if such an (internally represented) goal will be preferred to others– he will be regulated by my goal; for some motive he will act in order to realize my goal.

A very important case of goal-adoption (relevant for trust theory) is *goal–adhesion*, where *X* wants and expects that *Y* adopts her goal, communicates (implicitly or explicitly) this expectation or request to *Y*; *Y* knows that *X* has such an expectation and adopts *X*'s goal not unilaterally and spontaneously, but also because *X* wants it to be so. Thus not only does *Y* adopt *X*'s goal that *p*, but he also adopts *X*'s goal: 'that *Y* adopts her goal p'. In social trust frequently *Y*'s adoption (cooperation) is precisely due to *X*'s expectation and trust in *Y*'s adoption; and *X* relies on this response and adhesion.

We agree with Hardin ((Hardin, 2002); Chapter 1) that there is a restrict notion of social trust which is based on *the expectation of adoption* (or even *adhesion*), not just on the prediction of a favorable behavior of *Y*. When *X* trusts *Y* in the strict social sense and counts on him, she expects that *Y* will *adopt* her goal and this goal will prevail – in case of conflict with other

active goals. That is, *X* not only expects an *adoptive goal* by *Y* but an *adoptive decision and intention*. A simple *regularity* based prediction or an expectation simply based on some role or norm prescribing behavior to *Y*, are not enough – we agree with Hardin – for characterizing what he calls 'trust in strong sense', the 'central nature of trust', what we call 'genuine social trust'.

However, in our view, Hardin is not able to realize the broad theory of goal-adoption, and provides us – with his notion of *encapsulated interests* – a restricted and reductive view of it.

The various authors searching for a socially focused and more strict notion of trust go in this direction, but using a non general and not well defined notion, like: *benevolence, goodwill, other-regarding attitude, benignity* (Hart, 1988), *altruism, social-preferences, reciprocity*, *participant stance* (Holton, 1994).

And even the strange and unacceptable notion proposed by Deutsch (Deutsch, 1985) (we discuss this in Chapter 1 and Chapters 8) and repeated several times (for example, Bernard Barber 'to place the others' interests before their own') where in order to trust *Y* one should assume that he is altruistic or even irrational.

What *X* has to believe about *Y* is that:

i) *Y* has *some* motive for *adopting X*'s goal (for doing that action *for X*; for taking care of *X*'s interest); and that he will actually adopt the goal.
ii) Not only *Y* will adopt *X*'s goal (that is, he will formulate in his mind the goal of *X*, because it is the goal of *X*) but also that this goal will become an *intention*, so that *Y* will actually do as desired.

If (i) and (ii) are both true we can say that the adopted goal will prevail against other possible active goals of *Y*, including non-adopted goals (selfish).

More precisely we can claim that the motives *X* ascribes to *Y* while adopting *X*'s goal are assumed to prevail on the other possible motives (goals) of *Y*. Thus, what *X* is really relying on in genuine trust, are *Y's motives for an adoptive intention.*

The fact that a genuine social trust is based/relies on *Y*'s adoption should not be misinterpreted. One should not confuse *goal-adoption* with *specific motives* for adopting. Claiming that *X* counts on *Y*'s adoptive intention is not to claim that she counts on *Y*'s altruism, benevolence, good will, social preferences, respect, reciprocity, or moral norms. These are just specific sub-cases of the reasons and motives *Y* is supposed to adopt *X*'s goal. *X* might count on *Y*'s willingness to be well reputed (for future exchanges), or on his desire to receive gratitude or approval, or of avoiding blame or sanctions, or for his own approval, etc. In other words: *Y can be fully self-interested*.

To realize this it is necessary to keep in mind that the usual structures of goals are means-end chains: not all goals are *final goals*; they can be *instrumental* goals, simple means for higher goals. Thus, on the top of an adoptive and adopted goal there can be other goals, which *motivate* the goal-adoption. For example, I can do something *for* you, just in order to receive what I want for me, what you promise me.

'*It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages*' (Smith, 1776); however, when I ask the brewer to send me a box of beer and I send the money, I definitely *trust* him to give me the beer.

As seen in Section 1.5.7, we have three kinds of *social goal-adoption* (Conte and Castelfranchi, 1995): *Instrumental*, *Cooperative* and *Terminal*.

X can trust *Y*, and trusts that *Y* will do as expected, *for any kind of adoption*, also (or better, usually) *instrumental* (with both external or internal incentives). Trust in *Y* doesn't presuppose that *Y* is '*generous'* or that he will make '*sacrifices'* for *X*; he can strictly be selfish.

Now, we can formulate in a more reasonable way Deutsch's claim and definition, without giving the impression of trust as counting on *Y*'s altruism or even irrationality.

Y can be self-motivated or interested (autonomous, guided by his own goals) and can even be selfish or egoistic; what matters is that the intention to adopt *X*'s goal (and thus the adopted goal and the consequent intention to do α) will prevail on other non-adoptive, private (and perhaps selfish) goals of *Y*. But this only means that:

> *Y's (selfish) motives for adopting X's goal will prevail on Y's (selfish) motives for not doing so and giving precedence to other goals.*

So, *X* can count on *Y* doing as expected, in *X*'s interest (and perhaps for *Y*'s interest). Trustworthiness is a social 'virtue' but not necessarily an altruistic one. This also makes it clear that *not all 'genuine' trust is 'normative'* (based on norms) (for example, the generous impulse of helping somebody who is in serious danger is not motivated by the respect of a moral/social norm, even if this behavior (later) is socially/morally approved).

Moreover, *not all 'normative' trust is 'genuine'*. We can trust somebody for doing (or not doing α) just because we know that he has to do so (for a specific law or role), independently on his realizing or not and *adopting* or not our goal. For example, I trust a policeman for blocking and arresting some guy who was being aggressive to me, not because he has to respond to my desire, but just because he is a policeman at the scene of a crime (he can even ignore me).[51]

In sum, in genuine trust X just counts upon the fact that Y will understand and care of her (delegated) goal, Y will adopt her goal and possibly prefer it against conflicting goal (for example selfish ones), and this for whatever reason: from selfish advantages to altruism, from duty and obligations to cooperation, from love to identification, and so on.

In addition, May Tuomela (Tuomela, 2003) introduces and defines an interesting notion of 'genuine' social trust. But in our view this notion is too limited and specific. We disagree with her constraint that there is *genuine* trust only when it is symmetrical and reciprocal (for us this is counterintuitive and restrictive). In addition, her conditions (to be respected, the fact that the other will care about my rights, etc.) look quite peculiar in terms of specific – important – social relationships where there is 'genuine' trust, but which exclude other typical situations of trust (like child-mother) that must be covered.[52]

---

[51] It is also important to not mix up 'genuine' adoption-based trust with trust in 'strong delegation': delegation based on *X*'s request and *Y*'s acceptance. 'Genuine' trust can also be there in weak and in mild delegation/reliance: when *Y* ignores *X*'s reliance and acts on his own account, or when *Y*'s behavior is elicited by *X* (but without *Y*'s understanding). In fact, *Y* might have spontaneous reasons for adopting *X*'s interests (and *X* might count on and exploit this), or *X* might elicit in *Y* adoptive motives and attitudes by manipulating *Y*, without *Y* knowing that *X* is expecting and counting upon his adoption.

[52] Paradoxically, sometimes we trust Y precisely for his selfishness, which makes him trustworthy and reliable for that task/mission.

2.11.1 Concern

A very important notion in goal-adoption is the notion of *concern*. How much the goal of X is important for Y; how much Y is concerned with/by X's interest. That is, which is for Y the *value* of X's goal g_X, or best way of X achieving her goal. This value is determined by:

(i) the reasons (higher motivations) that Y has for adopting X's goal, and their value for him; how much and why Y cares about X's welfare;
(ii) $X's$ opinion about the subjective value of g_X for Y.

It is precisely on this basis that the adopted goal will prevail or not against possible costs, against other private conflicting goals of Y, and thus will possibly become/produce an adoptive *intention* of Y; and will also – as intention – persist against possible new interferences and temptations.

It is precisely on Y's *concern* for X's goal (not be confused with benevolence, good will, benignity, and so on) that X relies while betting on Y's adoptive intention and persistence. She also has some 'theory' about the reasons Y should be concerned with her welfare and wish to adopt her goal.

2.11.2 How Expectations Generate (Entitled) Prescriptions: Towards 'Betrayal'

It is characteristic of the most typical/genuine forms of social trust that – in case of failure – X is not only surprised and disappointed (Miceli and Castefranchi, 2002; Castelfranchi and Giardini, 2003), but feels *betrayed*. Where does this affective reaction come from? On which beliefs and goals (present in the trust attitude) is it based?

Social expectation can be entitled, can be based on Y's 'commitment' and thus obligation towards X (Castelfranchi, 1995). What X expects from Y can be 'due'. The violation of this kind of expectations involves not only disappointment but stronger and social emotions, like anger, indignation, etc. In particular it is different if this entitlement, this duty of Y towards X comes from legal norms or from interpersonal relations and merely social norms, like in a promise or like in friendship where fairness and adoption are presupposed.

In these forms of 'genuine' trust, where the expectation of Y's adopting/caring of my needs, requests, wishes (goals), is based on an assumption of *a moral duty* towards me, if Y disappoints this expectation I feel *betrayed* by Y in my trust and reliance on him.

This commitment – and the consequent moral duty, social norm –is not necessarily established in an explicit way; for example by a promise. Not only – as we said – can it be presupposed in the very relationship between us: friends, same family, same group, shared identity (which create some solidarity). It can be established by tacit consent, implicit behavioral communication (Castelfranchi, 2006; Tummolini and Castelfranchi, 2006). See Table 2.5 for an example.

In general, this mechanism is responsible for the tendency of shared social *expectations* (expectations about the behavior of the other agents, which are common knowledge) to become *prescriptions*: not only I predict that you will do something, but I wish so; I want you to behave in such a way (expectation). Moreover, I know that you know (etc.), and you did not disconfirm this (etc.), so you get some obligation of not violating my expectations. And I

Table 2.5 Example of Tacit Consent

IF	*X* decides to count on *Y*, AND *Y* is aware of such an expectation, AND *X* is aware that *Y* is aware, and *Y* knows this, AND *Y* would be able and in condition of rejecting such a delegation, to refuse to 'help' *X*, and to inform *X* about this (a very relevant information for her), and *Y* knows that *X* knows this; AND *Y* says nothing; doesn't provide any sign of his refusal
THEN	*Y* 'tacitly consent': *Y* takes a commitment, a tacit obligation to do as expected, to not disappoint *X*; AND *X* gets a *soft* right towards *Y*: she is entitled to ask and claim for *Y*'s action, and to complain and protest for his not doing as 'committed'

want you to do as expected also for this very reason: because you have a duty and you know (recognize) that I want you to do this simply for this reason. So my expectation becomes a true *prescription* (Castelfranchi and Giardini, 2003). This is how common expectations become social 'conventions' and 'norms'.

In sum, also on the bases of such tacit 'promises' and interpersonal norms, or of those obligations implicit in the relationship, *X* can feel betrayed by *Y*, since she was trusting *Y* on such a specific basis.

2.11.3 Super-Trust or Tutorial Trust

There are very extreme forms of trust, where *X* 'puts herself in *Y*'s hands' in a radical sense; in the sense that she believes and accepts that *Y* will care about her welfare better than her, beyond what she intends, asks, desires. One case is *over-trust*: trust in *Y*'s *over-help*. As we saw, we might confidently expect that, if needed, *Y* will do more than requested or will even violate our request in order to realize the adopted goal of ours. However, we can even go

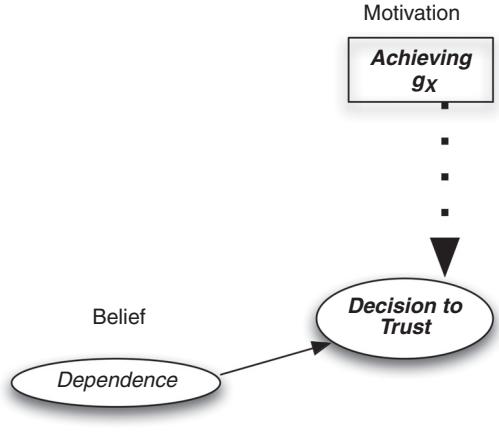

Figure 2.14 Dependence Belief and Goal: First step towards the decision of trusting

beyond this. In over-help *Y* is supposed to take care of our actual goals, desires, to seek out what we want; but there are forms of trust where we accept that *Y* goes against our current desires and objectives, while pursuing our (non understood) interests. This is *super-trust* or *tutorial trust*: trust in the 'tutorial' role of *Y* towards me. I feel so confident in *Y* that I am convinced that *Y* is pursuing my good, and is helpful, even when *Y* is acting against my current goals and I do not understand how he is taking care of me.

In other words, I assume that *Y* is taking care of my wellness, of doing the best for me, of my (possibly not so clear to me) interests, not just of my actual and explicit goals and desires (and may be against them). He does that for my good.

This presupposes that I feel/believe that I ignore part of my *interests*, of what is good for me (now or in the future), and I assume that, on the contrary, *Y* is *able* to understand better than me what is good for me (my interests) and *cares* about this, and *wants* – even against me – to protect my interests or oblige me to realize them.

We have modeled (Conte and Castelfranchi, 1995) this kind of social relationship between *Y* and *X* (when *Y* claims to know better than *X* what is better for *X*, and care about this, and

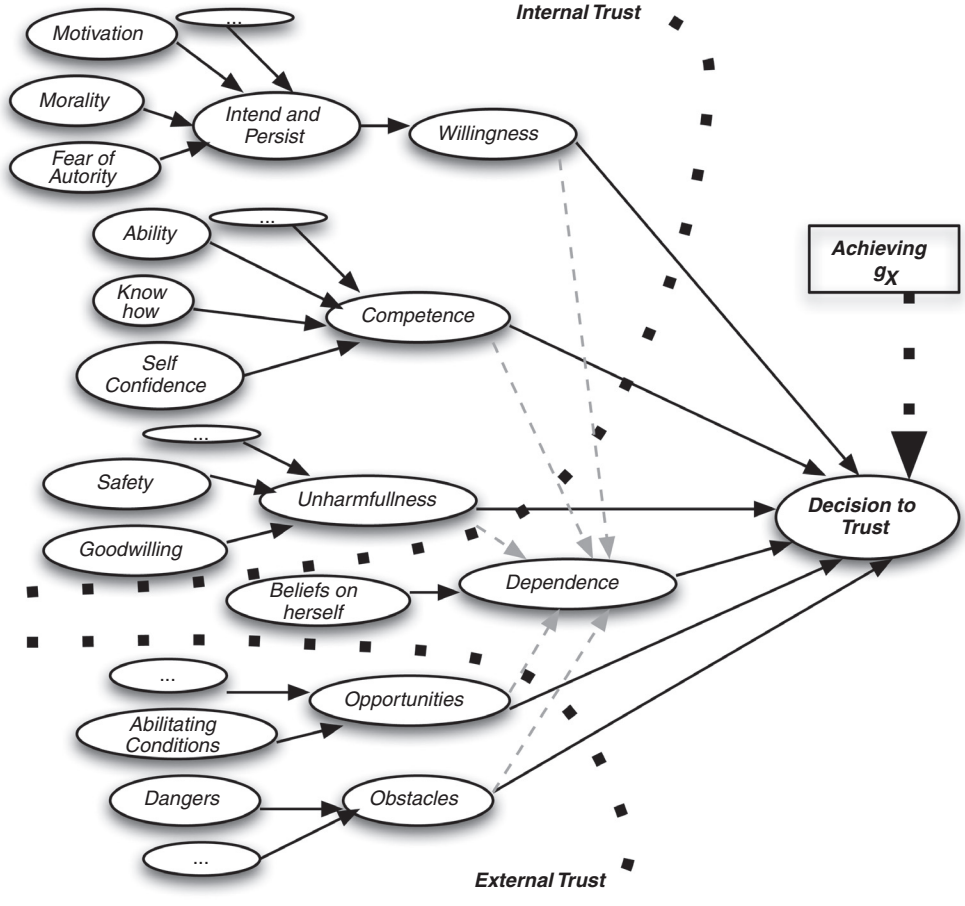

Figure 2.15 The complex set of beliefs converging towards the decision of trusting

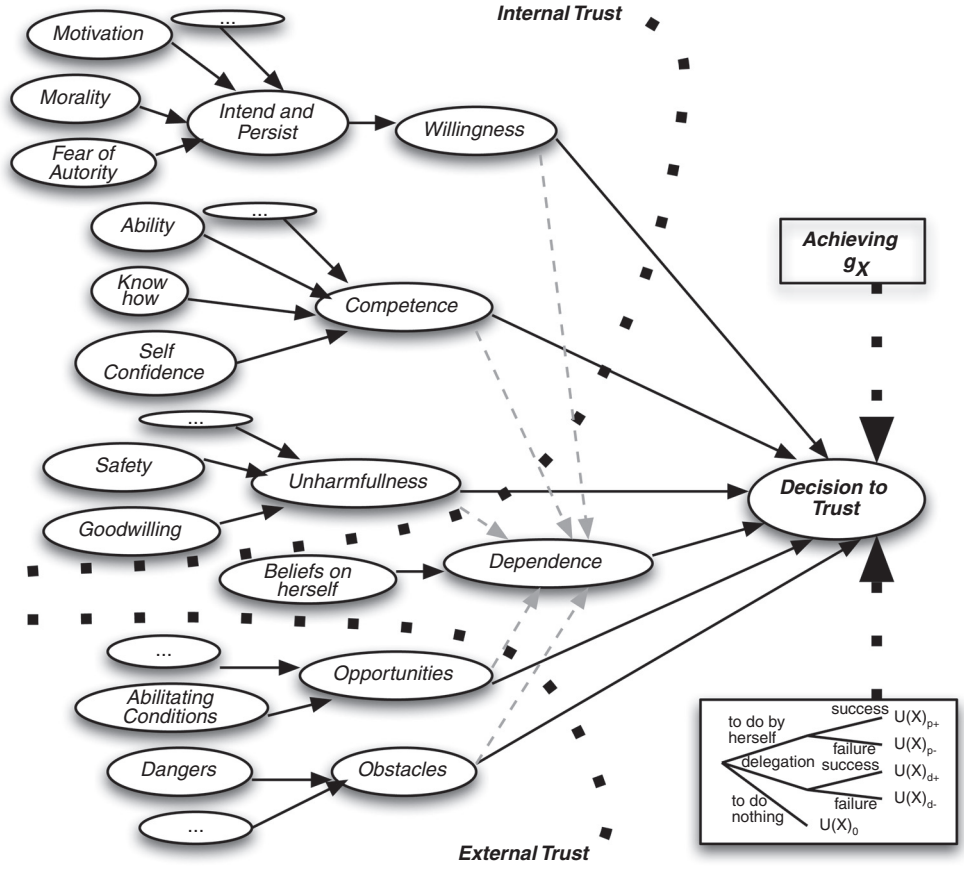

Figure 2.16 The role of the quantification in the complex scenario of Decision to Trust

try to influence X to do what is better for X); and we have labeled this *tutorial* relation of Y towards X. It also exists when X doesn't recognize or even contests it (like between parents and adolescents, or between psychiatrists and patients, etc.).[53]

In *super-trust*, X presumes a *tutorial* attitude and relation from Y, and relies on this, since he feels/believes that Y is really capable of understanding and will care about what is better for X.

2.12 Resuming the Model

Let us resume in a schematic and synthetic way how in our cognitive model of trust the different elements, playing a role in the trust concept, are composed and ordered for producing the trusting behavior of an agent. As we have seen a main role is played by the goal of the trustor that has to be achieved through the trustee (without this motivational component there

[53] However, sometimes this is just an arrogant and arbitrary claim, hiding Y's power and advantages, or 'paternalism'.

is no trust). In fact, in addition to the goal, it is also necessary that the trustor believes himself to be (strongly or weakly) dependent from the trustee himself see Figure 2.14.

On the basis of the goal, of her (potential) dependence beliefs,[54] of her beliefs about the trustee attributes (internal trust), of her beliefs about the context in which the trustee performance will come, the trustor (potentially) arrives at the decision to trust or not (Figure 2.15).

As explained in Section 2.2.1, all these possible beliefs are not simply external bases and supports of X's trust in Y (reduced to the Willingness and Competence and Dependence beliefs, and to the Decision and Act), but they are possible internal sub-components and forms of trust, in a recursive trust-structure. The frame looks quite complicated and complex, but, in fact, it is only a potential frame: not all these sub-components (for example, the beliefs about X's morality, or fear of authority, or self-esteem) are necessarily and already there or explicitly represented.

Moreover, as we will see in detail in Chapter 3, a relevant role is played by the quantification of the different elements: the weight of the beliefs, the value of the goal, the potential utilities resulting from a delegation and so on (see Figure 2.16).

References

Bacharach, M. and Gambetta, D. (2000) Trust in signs. In K. Cook (ed.), *Trust and Social Structure*. New York: Russel Sage Foundation.

Bacharach, M. and Gambetta, D. (2001) Trust as type detection, in: C. Castelfranchi, Y.-H. Tan (eds.), *Trust and Deception in Virtual Societies*, Kluwer Academic Publishing, Dordrecht, The Netherlands, pp. 1–26.

Baier, A. Trust and Antitrust, *Ethics* 96: 231–260, 1986.

Bandura, A. (1986) *Social Foundations of Thought and Action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Butz, M.V. (2002) *Anticipatory Learning Classifier System*. Boston, MA: Kluwer Academic Publisher.

Butz, M.V. and Hoffman, J. Anticipations control behaviour: animal behavior in an anticipatory learning classifier system. *Adaptive Behavior*, 10: 75–96, 2002.

Castaldo, S. (2002) *Fiducia e relazioni di mercato*. Bologna: Il Mulino.

Castelfranchi, C. Social Commitment: from individual intentions to groups and organizations. In ICMAS'95 *First International Conference on Multi-Agent Systems*, AAAI-MIT Press, 1995, 41–49 (versione preliminare in *AAAI Workshop on 'AI and Theory of Group and Organization'*, Washington, DC, May 1993).

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

Castelfranchi, C. (1997) Representation and integration of multiple knowledge sources: issues and questions. In Cantoni, Di Gesu', Setti e Tegolo (eds), *Human & Machine Perception: Information Fusion*. Plenum Press.

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

Castelfranchi, C. Towards an agent ontology: autonomy, delegation, adaptivity. *AI*IA Notizie.* 11 (3), 1998; Special issue on 'Autonomous intelligent agents', Italian Association for Artificial Intelligence, Roma, 45–50.

Castelfranchi, C. (2000a) Again on agents' autonomy: a homage to Alan Turing. *ATAL 2000:* 339–342.

Castelfranchi, C. (2000) Affective appraisal vs cognitive evaluation in social emotions and interactions. In A. Paiva (ed.) *Affective Interactions. Towards a New Generation of Computer Interfaces*. Heidelberg, Springer, LNAI 1814, 76–106.

Castelfranchi, C. Through the agents' minds: cognitive mediators of social action. *Mind and Society.* Torino, Rosembergh, 2000, pp. 109–140.

[54] To be true, the dependence belief already implies some belief about *Y*'s skills or resources, that are useful for *X*'s goal.

Castelfranchi, C. The micro-macro constitution of power, protosociology, an international journal of interdisciplinary research double vol. 18–19, 2003, *Understanding the Social II – Philosophy of Sociality*, Edited by Raimo Tuomela, Gerhard Preyer, and Georg Peter.

Castelfranchi, C. Mind as an anticipatory device: for a theory of expectations. In 'Brain, vision, and artificial intelligence' Proceedings of the First international symposium, BVAI 2005, Naples, Italy, October 19–21, 2005. Lecture notes in computer science ISSN 0302–9743 vol. 3704, pp. 258–276.

Castelfranchi, C. SILENT AGENTS: From Observation to Tacit Communication. *IBERAMIA-SBIA* 2006: 98–107.

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. Falcone R. (1997) Delegation Conflicts, Proceedings of the 8th European Workshop on Modelling Autonomous Agents in a Multi-Agent World: Multi-Agent Rationality, LNAI, Vol. 1237, pp. 234–254.

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., Giardini, F., Lorini, E., Tummolini, L. (2003) The prescriptive destiny of predictive attitudes: from expectations to norms via conventions, in R. Alterman, D. Kirsh (eds) *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston, MA.

Castelfranchi, C. and Lorini, E. (2003) Cognitive anatomy and functions of expectations. In *Proceedings of IJCAI'03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*, Acapulco, Mexico, August 9–11, 2003.

Castelfranchi, C. and Paglieri, F. (2007) The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese*, 155, 237–263.

Castelfranchi, C., Tummolini, L. and Pezzulo, G. (2005) From reaction to goals - AAAI workshop on From Reaction to Anticipation.

Cohen, J. (1992) *An Essay on Belief and Acceptance*, Oxford University Press, Oxford.

Conte, R. and Castelfranchi, C. (1995) *Cognitive and Social Action*. London: UCL Press.

Cook (ed.) *Trust and Social Structure*, New York: Russel Sage Foundation, forthcoming.

Cooper, J. and Fazio, R.H. A new look at dissonance theory. *Advances in Experimental Social Psychology*, 17: 229–265, 1984.

Dennett, D.C. (1989) *The Intentional Stance*. Cambridge, MA: MIT Press.

Deutsch, M. (1985) *The Resolution of Conflict: Constructive and destructive processes*. New Haven, CT: Yale University Press.

Drescher, G. (1991) *Made-up Minds: a constructivist approach to artificial intelligence*. Cambridge, MA: MIT Press.

Elster, J. (1979) *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

Engel, P. Believing, holding true, and accepting, *Philosophical Explorations* 1, 140–151, 1998.

Falcone, R. and Castelfranchi, C. (1999) Founding agents adjustable autonomy on delegation rheory, *Atti delSesto Congresso dell'Associazione Italiana di Intelligenza Artificiale (AI*IA-99)*, Bologna, 14–17 settembre, pp. 1–10.

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., Singh, M., Tan, Y.H. (eds.) (2001) Trust in cyber-societies. Lecture Notes on Artificial Intelligence, n°2246, Springer.

Falcone, R., Singh, M., and Tan, Y. (2001) Bringing together humans and artificial agents in cyber-societies: a new field of trust research; in *Trust in Cyber-societies: Integrating the Human and Artificial Perspectives* R. Falcone, M. Singh, and Y. Tan (eds.), LNAI 2246 Springer. pp. 1– 7.

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., Barber, S., Korba, L., Singh, M. (eds.) (2003) Trust reputation, and security: theories and practice. Lecture Notes on Artificial Intelligence, n°2631, Springer.

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

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

Good, D. (2000) Individuals, interpersonal relations, and trust. In Gambetta, D. (ed.) *Trust: Making and Breaking Cooperative Relations*, electronic edition, Department of Sociology, University of Oxford, pp. vii-x, <http://www.sociology.ox.ac.uk/papers/gambettavii-x.doc>.

Grosz, B. and Kraus, S. Collaborative plans for complex group action, *Artificial Intelligence*, 86:269–358.

Hardin, R. (2002) *Trust and Trustworthiness*, New York: Russel Sage Foundation.

Hart, K. (1988) Kinship, contract and trust: economic organization of migrants in an African city slum. In *Trust: Making and Breaking Cooperative Relations*, edited by Diego Gambetta. Oxford: Blackwell.

Hertzberg, L. (1988) On the Attitude of Trust. *Inquiry* 31 (3): 307–322.

Holton, R. Deciding to trust, coming to believe. *Australian Journal of Philosophy* 72: 63–76, 1994.

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

Jones, A. J. On the concept of trust, *Decision Support Systems*, 33 (3): 225-232, 2002, Special issue: *Formal modeling and electronic commerce*.

Johnson-Laird, P. N. (1983) *Mental models: towards a cognitive science of language, inference and consciousness*. Cambridge, UK: Cambridge University Press.

Levesque, H. J. (1984) A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence (AAAI-84)*, pages 198–202, Austin, TX.

Lorini, E. and Castelfranchi, C. (2007) The cognitive structure of Surprise: looking for basic principles. Topoi: An International Review of Philosophy, 26(1), 133–149.

Luhmann, N. (1979) *Trust and Power*, Wiley, New York.

Marsh, S.P. (1994) Formalising trust as a computational concept. PhD thesis, University of Stirling. Available at: http://www.nr.no/abie/papers/TR133.pdf.

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

McKnight, D. H. and Chervany, N. L. Trust and distrust definitions: one bite at a time. In *Trust in Cyber-societies*, Volume 2246 of Lecture Notes in Computer Science, pages 27–54, Springer, 2001.

Meyer, J.J. Ch. and Van Der Hoek, W. (1992) A modal logic for non monotonic reasoning. In W. van der Hoek, J.J. Ch. Meyer, Y. H. Tan and C. Witteveen, editors, *Non-Monotonic Reasoning and Partial Semantics*, pages 37–77. Ellis Horwood, Chichester, 1992.

Miceli, M., Castelfranchi, C. and Parisi, D. Verso una etnosemantica delle funzioni e dei funzionamenti: 'Rotto', 'guasto', 'non funziona', 'malato'. *Quaderni di semantica*, 8 (IV), 1983, 179–208.

Miceli, M. and Castelfranchi, C. (1997) Basic principles of psychic suffering: A preliminary account. *Theory & Psychology*, 7, 769–798.

Miceli, M. and Castelfranchi, C. (2000) The role of evaluation in cognition and social interaction. In K. Dautenhahn (ed.), *Human Cognition and Agent Technology* (pp. 225–261). Amsterdam: Benjamins.

Miceli, M. and Castelfranchi, C. (2002) The mind and the future: The (negative) power of expectations. *Theory & Psychology*. 12 (3): 335–366.

Miceli, M. and Castelfranchi, C. Anxiety as an 'epistemic' emotion: an uncertainty theory of anxiety. *Anxiety, Stress, and Coping*, 18: 291–319.

Miceli, M. and Castelfranchi, C. *Hope: the power of wish and possibility. Thoery & Psychology*, in press.

Miller, G. A., Galanter, E., and Pribram, K. A. (1960) *Plans and the Structure of Behavior*. New York: Holt, Rhinehart, & Winston.

Pezzulo, G., Butz, M., Castelfranchi, C. and Falcone, R. (Editors) *The Challenge of Anticipation*. LNAI State-of-the-Art Survey N°5225, pp. 3–22.

Searle, John R. (1995) *The Construction of Social Reality*, Free Press: NY.

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

Smith, A. (1776) *An Inquiry into the Nature and Causes of the Wealth of Nations*, London: Methuen & Co., Ltd.

Snijders, C. (1996) Determinants of Trust, *Proceedings of the workshop in honor of Amnon Rapoport, University of North Carolina at Chapel Hill*, USA, 6–7 August.

Tummolini, L. and Castelfranchi, C. Trace Signals: The Meanings of Stigmergy. E4MAS 2006: 141–156, 2006.

Tuomela, M. (2003) A collective's rational trust in a collective's action. In *Understanding the Social II: Philosophy of Sociality, Protosociology*. 18–19: 87–126.

van Linder, B. (1996) Modal Logics for Rational Agents, PhD thesis, Department of Computing Science, University of Utrecht.