1

Definitions of Trust: From Conceptual Components to the General Core

In this chapter we will present a thorough review of the predominant definitions of trust in the literature, with the purpose of showing that, in cognitive and social sciences, there is not yet a shared or prevailing, and clear and convincing notion of *trust*. Not surprisingly, this appalling situation has engendered frequent and diffuse complaints.[1] However, the fact that the use of the term *trust* and its analytical definition are confused and often inaccurate should not become an unconscious alibi, a justification for abusing this notion, applying it in any *ad hoc* way, without trying to understand if, beyond the various specific uses and limited definitions, *there is some common deep meaning, a conceptual core to be enlightened*.

On the contrary, most authors working on trust provide their own definition, which frequently is not really general but rather tailored for a specific domain (commerce, politics, technology, organization, security, etc.). Moreover, even definitions aimed at being general and endowed with some cross-domain validity are usually incomplete or redundant: either they miss or leave implicit and give for presupposed some important components of trust, or they attribute to the general notion something that is just accidental and domain-specific.

The consequence is that there is very little overlapping among the numerous definitions of trust, while a strong common conceptual kernel for characterizing the general notion has yet to emerge. So far the literature offers only partial convergences and 'family resemblances' among different definitions, i.e. some features and terms may be common to a subset of definitions but not to other subsets.

This book aims to counteract such a pernicious tendency, and tries to provide *a general, abstract, and domain-independent notion and model of trust*.

[1] See for example Mutti (1987: 224): 'the number of meanings attributed to the idea of trust in social analysis is disconcerting. Certainly this deplorable state of things is the product of a general *theoretical negligence*. It is almost as if, due to some strange self-reflecting mechanism, social science has ended up losing its own trust in the possibility of considering trust in a significant way'.

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

This theoretical framework:

- should take inspiration from and further analyze the common-sense notion of trust (as captured by natural languages), as well as the intuitive notions frequently used in the social sciences, but
- should also define a technical scientific construct for a precise characterization of trust in cognitive and social theory, while at the same time
- accounting for precise relationships with the most important current definitions of trust, in order to show what they all have in common, regardless of their different terminological formulations.

We believe this generalization and systematization to be both possible and necessary. In this chapter, we will start identifying the most recurrent and important features in trust definitions, to describe them and explain their hidden connections and gaps. This will be instrumental to a twofold purpose: on the one hand, we will show how our layered definition and quite sophisticated model can account for those features of trust that appear to be most fundamental; on the other hand, we will discuss why other aspects of current definitions of trust are just local, i.e. relevant only for a very specific problem or within a restricted domain. In this analysis, we will take as initial inspiration Castaldo's content analysis of trust definitions (Castaldo, 2002).

This critical effort will serve both to clarify the distinctive features of our own perspective on trust, and to highlight the most serious limitations of dominant current approaches.

1.1 A Content Analysis

In dealing with the current 'theoretical negligence' and conceptual confusion in trust definitions, Castaldo (Castaldo, 2002) applied a more descriptive and empirical approach, rather different but partially complementary to our own. Castaldo performed a content analysis of 72 definitions of trust (818 terms; 273 different terms), as employed in the following domains: Management (46%), Marketing (24%), Psychology (18%), and Sociology (12%). The survey covered the period from the 1960s to the 1990s, as described in Table 1.1:

Table 1.1 Number of trust definitions in different periods

Year	Definitions	Fraction
1960–69	4	(5.6%)
1970–79	5	(7.0%)
1980–89	19	(26.4%)
1990–99	44	(51.0%)
Total	72	(100.0%)

This table is from Castaldo. For more sophisticated data and comments, based on cluster analysis, see (Castaldo, 2002).

Source: Reproduced with kind permission of © 2002 Società editrice il Mulino.

This analysis is indeed quite useful, since it immediately reveals the degree of confusion and ambiguity that plagues current definitions of trust. Moreover, it also provides a concrete framework to identify empirically different 'families' of definitions, important conceptual nucleuses, necessary components, and recurring terms. Thus we will use these precious results as a first basis for comparison and a source of inspiration, and only later will we discuss in detail specific definitions and models of trust.

Castaldo summarizes the results of his analysis underlining how the trust definitions are based on five inter-related categories. They are:

- The *construct*, where trust is conceived 'as an *expectation*, a *belief*, *willingness*, and an *attitude*' (Castaldo, 2002).
- The *trustee*, 'usually individuals, groups, firms, organizations, sellers, and so on' (Castaldo, 2002). Given the different nature of the trustee (individuals, organizations, and social institutions), there are different types of trust (personal, inter-organizational and institutional). These trustees 'are often described by reference to different characteristics in the definitions being analyzed – specific competencies, capacities, non-opportunistic motivations, personal values, the propensity to trust others, and so on' (Castaldo, 2002).
- *Actions* and *behaviors*, as underlined also from other authors (e.g. (Moorman Zaltman and Desphande, 1992)) the behavioral aspect of trust is fundamental for 'recognizing the concept of trust itself' (Castaldo, 2002); both trustor and trustee behaviors have to take into account the consistence of the trust relationship. Behavioral aspects of trust have been studied also showing its multi-dimensional nature (e.g (Cummings and Bromiley, 1996)).
- *Results* and *outputs* of behavior, trustee's actions are presumed to be both predictable and positive for the trustor. 'The predictability of the other person's behavior and the fact that the behavior produces outcomes that are favorable to the trustor's objectives, are two typical results of trust. This has been particularly studied in works which suggest models designed to identify the consequences of trust (e.g. (Busacca and Castaldo, 2002)) (Castaldo, 2002).
- The *risk*, without uncertainty and risk there is no trust. The trustor has to believe this. They have to willingly put themselves into a 'position of vulnerability with regard to the trustee'. Risk, uncertainty and ambiguity (e.g. (Johannisson, 2001)) are the fundamental analytic presuppositions of trust, or rather the elements that describe the situations where trust has some importance for predictive purposes. (. . .).

[There is some sort of] logical sequence (. . .) [which has] often been suggested in the definitions. This sequence often regards trust as the *expectation, belief* (and so on) that a subject with specific characteristics (honesty, benevolence, competencies, and so on) *will perform* actions designed to produce *positive results* in the future for the trustor, in situations of consistent *perceived risk* (Castaldo, 2002).

Notwithstanding its merits, the main limit of Castaldo's analysis is that it fails to provide a stronger account of the *relationships* among these recurrent terms in trust definitions, i.e. indicating when they are partial synonyms, rather than necessary interdependent parts of a larger notion, or consequences of each other, and so on. Just an empirical, descriptive and co-relational account remains highly unsatisfactory. For example, it is true that 'Trust has been predominantly conceived as an *expectation*, a *belief*, *willingness*, and an *attitude*'.

However, it remains to be understood what are the conceptual ties between *belief* and *expectation*, or between *belief* and *willingness* (is one a species of the other? Does one

concept contain the other?). What are their exact roles in the processing of trust: For instance, what is the procedural relationship (e.g. sequential) between *belief* and *willingness*, which certainly is not a kind of belief? And why do some authors define trust *only* as a *belief*, while other authors only consider it as *willingness* and as a *decision* or *action*? Statistical relations do not even begin to address these questions.

An in-depth analysis of the *conceptual interconnections* among different facets of trust is also instrumental to achieve *a more adequate characterization of this notion*, since a good definition should be able to cover these different aspects and account for their relevance and their mutual relationships, or motivate their exclusion.

In particular, any theoretical model should take into account that trust is a *relational* construct, involving at the same time:

- A subject ***X*** (*the trustor*) which necessarily is an 'intentional entity', i.e. a system that we interpret according to Dennett's intentional stance (Dennett, 1989), and that is thus considered a cognitive agent.
- An addressee ***Y*** (*the trustee*) that is an *agent* in the broader sense of this term (Castelfranchi, 1998), i.e. an entity capable of causing some effect as the outcome of its behavior.
- The causal process itself (*the act,* or *performance*) and its result; that is, an act α of ***Y*** possibly producing the desired outcome ***O***.

Moreover, we should also never forget that trust is a *layered notion*, used to refer to several different (although interrelated) meanings (see Chapter 2):

- in its basic sense, trust is just a mental and affective *attitude* or *disposition* towards ***Y***, involving two basic types of *beliefs*: *evaluations* and *expectations*;
- in its richer use, trust is a *decision* and *intention* based on that disposition;
- as well as the *act* of *relying* upon ***Y***'s expected behavior;
- and the consequent social *relation* established between ***X*** and ***Y***.

If we now apply this analysis to the results summarized in Table 1.2, we can make the following observations:

- As for the terms ***Will***, ***Expect***, ***Belief***, ***Outcome***, ***Attitude***, they match the relation we postulate quite closely: *will* refers to the future (as Castaldo emphasizes), thus it is also included in the notion of *expectation*, which in turn involves a specific kind of *belief*: in its minimal sense, an expectation is indeed a belief about the future (Miceli and Castelfranchi, 2002; Castelfranchi and Lorini, 2003; Castelfranchi, 2003). Moreover, the term *belief* implies a mental attitude, and we can say that trust as evaluation and expectation is an *attitude* towards the trustee and his action: the *outcome*, the events, the situation, the environment.
- As for the terms ***Action*** and ***Decision***, they refer to trust as the deciding process of *X* and the subsequent *Y*'s course of action; hence they are general, but only with reference to the second and richer meaning of trust discussed above (see also below and Chapter 2).
- As for the terms ***Expect***, ***Outcome***, ***Rely***, ***Positive***, ***Exploit***, and ***Fulfill***, again they are tightly intertwined according to our relational view of trust: the positive outcome of the trustee's action is expected, relied upon, and exploited to fulfill the trustor's objective. In short: *X has*

Table 1.2 Most frequently used terms in trust definitions[2]

Terms	Frequency
Subject (Actor, Agent, Another, Company, Customer, Firm, Group, Individual, It, One, Other, Party, People, Person, Salesperson, Somebody, Trustee, Trustor)	180
Action (Action, Act, Behavior, Behave, Behavioral)	42
Will	29
Expect, Expectation, Expected, Expectancy	24
Belief, Believe	23
Outcome, Result, Performance, Perform	19
Rely, Reliable, Reliance, Reied, Reliability, Relying	18
Trust, Trusting, Trustworthy	17
Confident, Confidence	16
Willingness, Willing	14
Take, Taken, Taking, Accept, Accepted, Acceptable	11
Risk, Risky, Risking	11
Vulnerable, Vulnerability	11
Relationship	10
Exchange	9
Based	8
Competent, Competence, Capabilities	7
Positive	7
Cooperate, Cooperation, Coordination	6
Exploit, Exploitation	6
Situation	6
Attitude	5
Decide, Decision	5
Fulfill, Fulfilled, Fulfillment	5
Held	5
Intention, Intentionally, Intend	5
Involve, Involved, Involvement, Involving	5
Mutual, Mutually	5
Word	5
Would	5

Source: Reproduced with kind permission of © 2002 Società editrice il Mulino.

a goal (a desire or need) that is expected to be fulfilled thanks to Y's act; X intends to exploit the positive outcome of Y's act, and relies upon Y for fulfilling the goal.

- As for the terms ***Taken***, ***Accept***, ***Risk***, and ***Vulnerable***, their relationship is that while deciding to count on *Y*, to trust *Y* (according to trust as decision), *X* is necessarily accepting the risk of becoming *vulnerable* by *Y*, since there is uncertainty both in *X*'s knowledge (incomplete, wrong, static) and in the (unpredictable, unknown) dynamics of the world.

[2]This table is from Castaldo. For more sophisticated data and comments, based on cluster analysis, see (Castaldo, 2002).

Whenever deciding to depend on Y for achieving O, X is exposed both to failure (not fulfilling O) and to additional harms, since there are intrinsic costs in the act of reliance, as well as retreats to possible alternatives, potential damages inflicted by Y while X is not defended, and so on. As we will discuss more thoroughly in the next chapters, all these risks are direct consequences of X's decision to trust Y.

- As for the terms ***Competence*** and ***Willingness***, they identify the two basic prototypical features of 'active'[3] *trust in Y*, i.e. the two necessary components of the positive evaluation of Y that qualify trust:

  - The belief of X (evaluation and expectation) that Y is *competent* (able, informed, expert, skilled) for effectively doing α and produce O;
  - The belief of X (evaluation and expectation) that Y is *willing* to do α, intends and is committed to do α – and notice that this is precisely what makes an agent Y predictable and reliable for X. Obviously this feature holds only when Y is a cognitive, intentional agent. It is in fact just a specification of a more abstract component that is Y's *predictability*: the belief that 'Y will actually do α and/or produce O', contrasted with merely having the potentiality for doing so.

In sum, a good definition of trust, and the related analytical model that supports it, must be able to explicitly account for two kinds of relationships between the different components of this multi-layered notion: *conceptual/logical links*, and *process/causal links*. A mere list of relevant features is not enough, not even when complemented with frequency patterns.

More specifically, a satisfactory definition should be able to answer the following questions:

1. What are the relevant connections between the overall phenomenon of trust and its specific ingredients? Why are the latter within the former, and how does the former emerge from the latter?
2. What are the pair-wise relations between different features of trust? For instance, how do *belief* and *expectation*, or *outcome* and *reliance*, interact with each other?
3. What is the conceptual link and the process relationship between trust as attitude (*belief*, *evaluation*, *expectation*) and trust as decision and action (relying on, accepting, making oneself vulnerable, depending, etc.)?

1.2 Missed Components and Obscure Links

The content analysis of 72 definitions presented in the previous section reveals some relevant gaps in such definitions, as well as several notions that remain largely or completely implicit.

An aspect absolutely necessary but frequently ignored (or at least left unstated) is *the goal, the need,* relative to which and for the achievement of which the trustor counts upon the trustee.

[3] As we will discuss later on (Chapter 2, Section 2.4), we distinguish between *active trust* and *passive trust*. The former is related to the delegation of a positive action to Y, and to the expectation of obtaining the desired outcome from this action. The latter, instead, is just reduced to the expectation of receiving no harm from Y, no aggression: it is the belief that Y will not do anything dangerous for me, hence I do not need to be alerted, to monitor Y's behavior, to avoid something, to protect myself. This passive trust has a third, more primitive component: the idea or feeling that "there is nothing to worry about", "I am/feel safe with Y".

This is implicit when a given definition of trust mentions some 'positive' result/outcome, or the 'welfare' or 'interest' of the trustor, and also whenever the notions of 'dependence', 'reliance', or 'vulnerability' are invoked. In fact, something can be 'positive' for an agent only when this agent has some *concern, need, desire, task,* or *intention* (more generally, a goal), because 'positive' means exactly that the event or state or action is favorable to or realizes such a goal – whereas 'negative' means the opposite, i.e. a threat or frustration of some goal.

Analogously, whenever it is observed that the trustor makes her/himself vulnerable to the trustee (see for instance (Mayer *et al.*, 1995)), the unavoidable question is – vulnerable for what? Alongside other costs that are intrinsic in any act of reliance, the trustor becomes especially vulnerable to the trustee in terms of potential failure of the expected action and result: the trustee may not perform the action α or the action may not have the *desired* result O.

Moreover, it is precisely with reference to the desired action/result that the trustor is 'dependent on' and relies upon the trustee. Also in the famous definitions provided by (Deutsch, 1985) where trust is relative to an entity 'on which my *welfare* depends', the goal of the trustor is clearly presupposed, since the notion of 'welfare' refers to the satisfaction of her needs and desires.

Building on these observations, in the following we will extensively argue for the necessity of the trustor to be actively *concerned*, i.e. to have goals at play in the decision to trust someone or something, as well as full *expectations*[4] rather than mere beliefs on the future (forecasts).

Another aspect frequently missed is that trust is an ***evaluation***, and more exactly a *positive evaluation about Y*. In the cognitive architecture developed by (Miceli and Castelfranchi, 2000; Castelfranchi, 2000), an evaluation (which is also an attitude) is a *belief about some power* (capacity, ability, aptitude, quality; or lack of capacity . . .) *of Y relative to some goal*. Thus the beliefs about Y being able and willing to do his share for achieving O are in fact evaluations of Y: positive when there is enough trust, negative when there is mistrust and diffidence.

Here it is important to appreciate the intimate relation between 'beliefs', 'expectations' and 'evaluations'. In the case of trust, the *beliefs* on the competence and willingness of Y are both, and at the same time, *parts of expectations* (since they are about the future) and *evaluations* (since they are about Y's powers and inclinations); moreover, they are *positive* both as expectations and as evaluations, insofar as agent X is expecting from and attributing to Y an attitude and a subsequent performance that is in X's best interest. In addition, X might ground the decision to trust also on other positive evaluations of Y, for example, intelligence, honesty, persistency (on this point, see also Chapter 2).

It is worth noticing that the characterization of *trust as a structure of mental dispositions* is not in contrast with the analysis of *trust as a decisional process culminating into an action* – quite the contrary. Indeed, it is rather pointless to dispute whether trust is a *belief* or an *act*, opposing the view of trust as an *evaluation* or *expectation* to the idea that trust is a *decision* or a *behavior* (for example, of making oneself vulnerable, of risking, of betting on Y). The point is rather that *trust has both these meanings*, which stand in a specific structural relation with each

[4] As detailed in Chapter 2 (see also (Castelfranchi, 2005)), by expectation we mean the functional integration of a belief on the future (forecast) with a goal, i.e. a motivational state. In short, an expectation (either positive or negative) is defined as the prediction that a state of the world which constitutes one of the agent's goals will either be realized or not in the future. Thus, both the goal that p and the belief that, at some time in the future, p will (or will not) be the case are necessary components for expecting that p (or *not-p*).

other: more precisely, trust as decision/action presupposes trust as evaluation/expectation, as we shall argue in Chapter 2.

Yet another issue that remains often underestimated or confused is the *behavioral aspect of trust*, i.e. all the different types of actions that distinct actors (roles) have to perform, in order for trust to be a felicitous choice.

As for ***the act of the trustee***, what is frequently missed is that:

(a) The act can be *non-intentional*. First, the trustee might not be aware of my reliance on him/it,[5] or he/it may not know or intend the specific result of its action that I intend to exploit (see for instance the famous anecdote of Kant and his neighbor, where the latter was relying on the former and trusting him for his punctuality and precision, without Kant being aware of such a reliance). Second, the act that I exploit can be a non-intentional act by definition: e.g. a reactive behavior, or a routine. Third, if we endorse a very general notion where one can also trust natural processes or artifacts (see (c)), then of course the exploited process and the expected result that we delegate to a natural event or to an artifact are not intentional.

(b) Also *omissions* may be relevant in this context: e.g. 'doing nothing', 'not doing α', 'abstaining from α'. Obviously, omissions can be acts, even intentional ones – in which case, they are the result of a decision. In addition, omissions can also be the outcome of some more procedural choice mechanism, or of a merely reactive process, as well as just the static and passive maintenance of a previous state (i.e. they are not even proper 'acts' in the latter sense). Regardless the specific nature of the omission, the trustor might precisely expect, desire, and rely on the fact that Y will not do the specific action α, or more generally that Y will not do anything at all (See note 3).

(c) The trustee is *not necessarily a cognitive system*, or an animated or autonomous agent. Trust can be about a lot of things we rely upon in our daily activity: rules, procedures, conventions, infrastructures, technology and artifacts in general, tools, authorities and institutions, environmental regularities, and so on. Reducing trust to 'trust *in* somebody' is not only an arbitrary self-limitation, but may also bias the proper interpretation of the phenomenon, insofar as it hides the fact that, even when we trust somebody's action, we are necessarily trusting also some external and/or environmental conditions and processes (on this point, see Chapter 2). However, it remains obviously very important to precisely characterize *social trust*, i.e. trust in another agent as an agent, and the so called 'genuine' or typical trust in another human (see Chapter 2).

As for **the act of the trustor**, the more frequent shortcomings and confusions in the literature are the following:

(a) It is often missing a clear (and basic) distinction between the act of the trustor and the act of the trustee: for instance, Castaldo does not clearly disentangle the occurrences of the two different acts within the definitions covered by his survey. Obviously, the act of the trustor consists in the very act of 'trusting', of counting upon and deciding to rely on Y.

(b) Much more importantly, it is not always emphasized enough that the 'act' of trusting is a necessary ingredient of one notion of trust (i.e. as a decision and a subsequent action), but

[5] "It" since it can even be a tool, an inanimate active entity (agent).

not of the other notion of trust as a preliminary mental attitude and evaluation. Since this crucial distinction is considered as a semantic ambiguity, rather than a valuable insight into the internal articulation of this complex phenomenon, it is still lacking the theory of the logical and causal relationships between these two aspects of trust: the mental attitude and the decision to act upon it (see Chapter 2).

1.3 Intentional Action and Lack of Controllability: Relying on What Is Beyond Our Power

In any intentional action α exerted upon the external world, there is one part of the causal process triggered by the action and necessary for producing its intended and defining result (the goal of the action) which *is beyond the direct executive control of the agent* of α. Whenever an agent is performing α in the real world, there is always some external condition or process P that must hold for the action to be successful, and the agent does not have direct executive control over such an external feature – although he might have foreseen, exploited, or even indirectly produced it. Therefore, the agent while performing α is objectively making reliance on these processes in order to successfully realize the whole action and thus achieve his goal.

This objective reliance holds in both of the following cases:

- when the agent is aware of this fact, models this act of reliance in his mind, and even expects it;
- when the agent does not understand the process, he is not aware of it, or at least he does not explicitly represent it in his plan (although in principle he might be able to do so).

When the subject is aware of the reliance that he is making for his action α on some external process, and counts upon such a process P, which does not depend completely and directly on him, we can say that the reliance has become *delegation*. Delegation (Castelfranchi and Falcone, 1998; Falcone and Castelfranchi, 2001) is a type of reliance, which is subjective (i.e. aware) and decided by the agent; it consists of the act of 'counting upon' something or someone, which is both a mental disposition and a practical conduct.[6]

In contrast, *reliance* in general can be merely objective or also subjective, e.g. like in delegation. When reliance is subjective, it can (like delegation) be correct or wrong and illusory: e.g., the beliefs on which it is based may be false, it may not be true that that expected process will be there or that it is responsible for the desired effect (like happens, for instance, with a placebo).

It is worth noticing that, although the presence of P (due to Y) is a necessary process/condition for the achievement of X's own goal, it is not sufficient. X has also *to do (or abstain from doing) something of his own*, and thus he has *to decide something*: regardless that X is counting on Y for P or not, he still has to take his own decision on whether to pursue

[6] Here we use 'delegation' in its broader and non-organizational meaning: not only and not necessarily as delegation of powers, or delegation of institutional/organizational tasks from one role to another (involving obligations, permissions, and other deontic notions). Our use of delegation is more basic, although strictly related to the other: to delegate here means *to allocate, in my mind and with reference to my own plan, a given action that is part of the plan to another agent*, and therefore *relying on the performance of such an action by the other agent* for the successful realization of my own plan.

the goal in the first place, i.e. on whether to engage in action α (that would make *X* dependent upon *Y* for *P*) or not.

The link between delegation and trust is deep and should not be underestimated: delegation necessarily requires some trust, and trust as decision and action is essentially about delegation. This also means that *trust implies that X has not complete power and control on the agent/process they are relying and counting upon*. Trust is a case of limited power, of 'dependence' on someone or something else. Although the notion of 'reliance' and 'reliable' is present in several of the definitions analyzed by Castaldo, the theory of this strange relation, and its *active* aspect of deciding to depend, deciding to count on, to invest, to delegate, is not yet well developed in the literature on trust. For instance, several authors consider a crucial and necessary aspect of trust the fact that while relying on the behavior of another person we take a risk because of the lack or limit of '*controllability'* and because the other's behavior *cannot be under coercion*, so that our expectation on the other's actions cannot be fully certain. This intuition is correct, and it just follows from our previous analysis. In any act of *trusting in Y* there is some *quid* delegated to another agent *Y*, and, especially when this agent is viewed as purposive, goal-oriented (be it Nature, a tool, an animal, or a person), the delegated process that consists of *Y*'s performance is *beyond our direct control*. *Y* has some autonomy, some *internal degree of freedom*, and also for this – not only for external interferences – it is not fully predictable and reliable.

When *Y* is an autonomous cognitive agent this *perceived* degree of freedom and autonomy mainly consists in *Y*'s choice: *Y* can *decide* to do or not to do the expected action. With this kind of agent (i.e. within the domain of social trust), we in fact trust *Y* for deciding and be willing to do what *Y* 'has to' do (for us) – even against possible conflicting goals that may arise for *Y* at the very moment of the expected performance. In other words, we trust (in) *Y*'s motivation, decision, and intention.

This characteristic feature of social trust strictly derives from the more basic notion of trust as involving reliance upon some non-directly controlled process and agent, on the perception of this lack of controllability, and its associated risk; on the distinction between trust 'in' *Y*, and global trust; and, in the end, on the very idea of 'delegation', as the decision to count upon such a process/agent for the pursuit of my own goal. If I have not decided to depend on this, I would have no reason to care for any non-controllability of such a process or agent.

Finally, as we briefly mentioned before, aside from *Competence* and *Willingness*, there is a third dimension in evaluating the trustworthiness of *Y*: *Y* must be perceived as non threatening (passive trust), i.e. *harmless*. There is no danger to be expected from *Y*'s side; it is 'safe' to rely on *Y* and to restrain from fully monitoring *Y*'s conduct (see Section 2.4 for more details).

In a sense 'feeling safe' could be taken as the basic nucleus of trust in and by itself, seemingly without any additional component (*passive trust* – see note 3). However, looking more carefully we can identify other core components. Clearly positive *evaluations* and *expectations* (beliefs) are there and play a role in my feeling of safety. If I do not worry and do not suspect any harm from you, this means that I evaluate you positively (in the sense that you are 'good for me', not to be avoided, at least harmless), since not being harmed is one of my goals. I can be relaxed (as for you) and this is also pleasant to me. Moreover, this feeling/belief is an expectation about you: I do not expect damage from you, and this constitutes a passive, weak form of positive expectation. Perhaps I do not expect that you will actively help me realize a goal of mine that I am pursuing; but at least I expect that you will not compromise a goal that I have

already achieved, i.e. you will not damage me, but rather let me continue to have what I have. (This is why we call 'passive' this kind of expectation).

One of the clusters of definitions found by Castaldo is the definition of trust as a 'Belief about future actions' (27, 8%) that 'makes no reference to the concept of Expectations'. This is in our interpretation either due to the fact that the authors adopt a very weak notion of 'expectation' and thus simply analyze it in terms of 'a belief about the future'; or, instead, they have a richer and stronger notion of 'expectation' but expunge this notion from the definition of trust, ultimately reducing trust to some *grounded prediction*. In both cases, this position is unsatisfactory; trust, even when considered as a pure mental attitude before and without any decision and act, cannot be reduced to a simple forecast (although it contains this element of prediction within its kernel). Computers, if adequately instructed, can make excellent weather forecasts, but they do not have 'expectations' about the weather, nor do they put their 'trust' in it – indeed, they remain incapable of trusting anything or anyone, as long as they are mere forecasting machines. In what follows, we will take Andrew Jones' analysis of trust (Jones, 2002) as a good example of this tendency to reduce trust to an epistemic attitude, to grounded prediction, and discuss it in order to show why such an analysis necessarily misses some very important trait of the notion of trust (see Section 1.4 and Chapter 2).

Some of the most frequent terms highlighted by Castaldo's content analysis, like ***Cooperate, Mutually, Exchange, Honesty, Commitment, Shared Values***, are clearly valid for describing trust only in specific domains and situations, e.g. commerce and organization. *Mutuality*, for instance, is not necessary at all: in most contexts, trust can be just unilateral – and in fact later on we will criticize this same bias in philosophical, economic and game theoretic theories of trust. Meanwhile, it is a real distortion of the game theoretic perspective to use 'trusting' as a synonym of 'cooperating' (see below and Chapter 8). Analogously, the terms ***Customer, Company, Salesperson, Firm*** (gathered by Castaldo in the category of 'Subjects') are clearly domain-specific. The same would hold for the term ***Security*** in the growing domain of 'Trust and Security' in Information and Communication Technologies (see Chapter 12).

1.4 Two Intertwined Notions of Trust: Trust as Attitude vs. Trust as Act

Although we are aiming for a unified, covering, general and possibly shared definition of trust, this result will not be achieved by looking for just one unique monolithic definition. We also do not want to gather just a list of different meanings and uses that share with each other only some features; we can accept this 'family resemblance' as a possible result of the conceptual analysis, but not as its starting assumption and its ultimate objective.

Ideally, what we will try to identify is a *kernel concept*: *few common and truly fundamental features*. In doing so, what is needed – as is often the case (see for instance Castelfranchi's definition of 'agent' in (Castelfranchi, 1998; Castelfranchi, 2000a; Castelfranchi and Falcone, 2003)) – is a *layered definition* or a hierarchy of definitions, with an explicit account of the conceptual relationships between the different but connected meanings of the same notion.

The common sense term *trust* (at least in many of the languages used by the scientific community, like English, German, French, Spanish, Italian) covers various phenomena structurally connected with each other. As we said, it is crucial to distinguish at least between two kinds

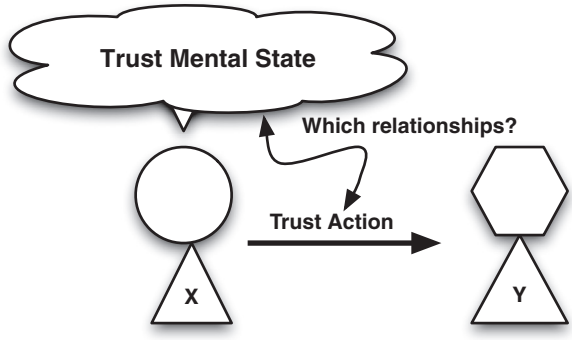

Figure 1.1 The double nature of Trust, as psychological attitude and as act. The relationship between these two elements needs much exploration

and meanings of trust (see Figure 1.1):

(a) Trust as *psychological attitude* of *X* towards *Y* relative to some possible desirable behavior or feature.
(b) Trust as the *decision and the act of relying on*, counting on, depending on *Y*.

In our theory there is a twofold connection between these two types of trust:

- The *conceptual link* in that the *intension* of *(b)* contains *(a)*, i.e. trust as an attitude is part of the *concept* of trust as a decision/action. Obviously, the *extension* of *(a)*, i.e. the set of cases where there is such a mental attitude, includes the extension of *(b)*, i.e. the set of acts of trusting someone or something.
- The *process/causal link* in that *(a)* is the temporal presupposition (an antecedent) and a con-cause of *(b)*. The process for arriving at the act of trusting entails the formation of such a positive expectation on and evaluation of *Y*.[7]

This provides us with the dyadic general notion of trust, capable of connecting together the two basic meanings of this phenomenon.[8] Revolving around this common core, there are then several domain-related specifications of trust: subcategories and/or situational forms of trust, due to the sub-specification of its arguments (the kind of trustor or of trustee; the kind of action; the kind of results), or to different contexts that make relevant some specific reason of trust (like interest, commitment, or contract) or some specific quality of the trustee (like sympathy, honesty, friendship, common values).

[7] Without the psychological attitude (a) there might be delegation but not trust: An obliged, constrained, needed, (may be) desperate delegation/reliance but not trust.

[8] One might more subtly distinguish (see (Falcone and Castelfranchi, 2001b)) between the mere decision to trust and the actual act of trusting. In this case the definition of trust becomes three-layered, involving a three-steps process (Chapter 2). One might also argue that there is trust as a social relation, in consequence of the act of trusting or of a trustful or distrustful attitude towards someone else. This is certainly true, so one might enrich a layered definition of trust along the lines suggested by these considerations.

1.5 A Critique of Some Significant Definitions of Trust

At the beginning of this chapter we introduced our aim to review the predominant definitions of trust in the literature, motivated by the purpose of showing that, in cognitive and social sciences, there is not yet a shared or prevailing, and clear and convincing notion of *trust*. After a content analysis of a larger number of definitions, let us now consider some specific definitions in order to show how they frequently are incomplete, vague or obscure, domain specific, divergent. To show this, we will be a bit simplistic and not completely fair with the authors, just discussing and criticizing their definitions of trust. This is useful for comparing them, and for stressing confusion, inaccuracy, and ad hoc features; but it is quite ungenerous, since sometimes the analysis or theory of the author is more rich and correct than their 'definition'. However, several authors are more extensively discussed here or in other chapters of this book.

1.5.1 Gambetta: Is Trust Only About Predictability?

Let us first consider the definition of trust provided in the classic book of Gambetta and accepted by the great majority of other authors (Gambetta, 1988): 'Trust is the *subjective probability* by which an individual, *A*, *expects* that another individual, *B*, performs a given action on which *its welfare depends*' (translation from Italian).

In our view, this definition is correct, insofar as it stresses that trust is basically an estimation, an opinion, an expectation, i.e. a belief. We also find commendable that there is no reference to exchange, cooperation, mutuality, and *B*'s awareness of being trusted by *A*, since none of these features is, in our view, part of the core notion of trust.

However, Gambetta's definition is also too restricted in a variety of respects:

- It just refers to one dimension of trust (*predictability*), while ignoring the *competence* dimension.
- It does not account for the meaning of '*A* trusts *B*' where there is also the *decision* to rely on *B*.
- It does not explain what such an evaluation is made of and based on, since the measure of *subjective probability* collapses together too many important parameters and beliefs, which each has its own role and relevance in social reasoning.
- It fails to make explicit the 'evaluative' character of trust.
- Finally, reducing trust to the notion of 'subjective probability' is quite risky, since it may result in making superfluous the very notion of 'trust' (on this point, see Williamson's criticism (Williamson, 1993) as well as Chapter 8).

1.5.2 Mayer, Davis, & Schoorman: Is Trust Only Willingness, for Any Kind of Vulnerability?

Mayer, Davis, and Schoorman provide an interesting and insightful (albeit somehow limited) definition of trust, as follows: 'The *willingness* of a party *to be vulnerable* to the actions of another party based on the *expectation* that the other party will perform a particular action

important to the trustor, irrespective of the ability to monitor or control that other party' (Mayer *et al.*, 1995).

This definition of trust as *the decision to make oneself vulnerable* deserves some consideration. On the one hand, it is strange that it focuses only on trust as a decision, without considering the background of such a decision, and thus missing basic uses of the term like in 'I trust John but not enough'.

On the other hand, in our view it identifies and expresses very well an important property of trust, but without providing a really good definition of the general phenomenon. To begin with, insofar as the definition is mainly applicable to trust as decision and action, it seems to allude to vulnerability only in relation to a transition of state, whereas one might also say that a disposition of trust or a relation of trust is enough to make the trustor vulnerable, although in a static sense, not as state-transition.

More importantly, the idea of equating trust with self-decided vulnerability is, as a definition, both too broad and too vague, since there are a lot of states (including psychological states) and acts that share the same property of making oneself vulnerable to others; for example, *lack of attention and concentration*, *excess of focus and single-mindedness*, *tiredness*, *wrong beliefs about dangers* (e.g. concerning exposition to an enemy, being hated, inferiority, etc.), and so on. Moreover, some of these states and acts can be due to a decision of the subject: for example, the decision to elicit envy, or to provoke someone. In all these cases, the subject is deciding to make themselves vulnerable to someone or something else, and yet no trust is involved at all.

Therefore, the problem is not solved by defining trust as the decision to make oneself vulnerable, although one should characterize trust in such a way that can *explain* and *predict* these important effects and consequences of trust. For instance, it is worth emphasizing all the dangers implicit in the decision to trust, in terms of:

- Considering sufficient and reliable enough the current information about the trustee and about the relevant situation. This implies that the trustor does not perceive too much uncertainty and ignorance on these matters, although their estimate is, of course, subjective and fallible (and ignorance or wrong certainty can be dangerous indeed).
- Having enough good evaluations and predictions about the trustee; but these might be wrong, or the negative evaluations and foreseen dangers can be poorly estimated, and false predictions based upon misleading evaluations may be extremely noxious.
- Relying upon Y, counting on Y to help realize a given goal; i.e. for the realization of goal G, agent X depends (accepts, decides to depend) on Y. This – analytically in our model– gives Y the power of frustrating X's goal, thus X makes oneself vulnerable to Y for G; moreover the actual decision to trust further increases X's dependence.

So, in order to explain and predict trust-related vulnerability and the fact that the trustor's welfare comes to depend upon the trustee's action (as mentioned in Gambetta's definition), a model of trust – as we said – must at least integrate:

- beliefs about the trustee's internal attitudes and future conduct (more or less complete; more or less grounded on evidence and rationally justified; more or less correct);
- the subjective propensity of the trustor to accept a given degree of uncertainty and of ignorance, and a given perceived amount of risk;

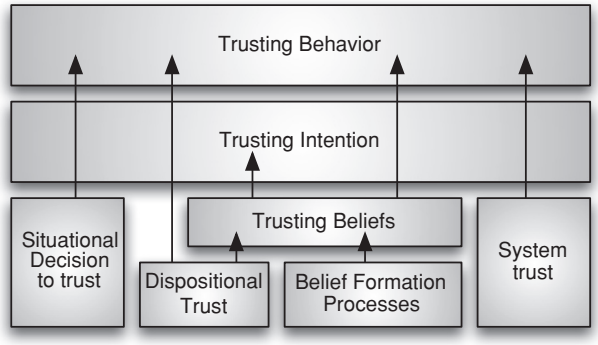

Figure 1.2 Relationships among Trust Constructs (arrows represent relationships and mediated relationships). (Reproduced with kind permission of Springer Science+Business Media)

- the trustor's decision to rely upon the action of another entity for the realization of a goal, and the expectations upon which such a decision is based;
- the relationships of dependence and power between the trustor and the trustee with respect to the intended goal of the former.

1.5.3 McKnight: The Black Boxes of Trust

A very apt and frequently cited approach to the nature and internal dynamics of trust is McKnight's model (McKnight and Chervany, 2001). In Figure 1.2 one can see the general schema of this model.

This model is rather comprehensive, since it takes into account several important aspects and some of their mutual interactions. For example, the authors are able to distinguish between the belief-component and the decisional and behavioral aspects of trust, and to explain that the latter depends on the former; they also recognize the role of situational and system components in determining trust behavior. However, this is just a black-boxes model, without much insight on what is supposed to be going on *within* each box. Moreover, the semantics of the arrows is undefined and definitely non-uniform. The specific nature, organization, structure and process of the content of the various boxes are not well specified. There is no deeper characterization of trust beliefs in terms of 'expectations' and 'evaluations', nor is there an explicit model of the critical factors involved in the decision process.

On the whole, this remains a 'factors' model (typical of psychology), where the authors just capture correlations and mutual influences, but the precise nature and 'mechanics' of the process are not defined. In sum, it is indeed important that they identify different kinds and levels of trust (as beliefs, as intention, as behavior), and that they connect each one with the others. However, a much more *analytic*, *process-oriented*, and *relational* (not only mental) model is needed.

1.5.4 Marsh: Is a Mere Expectation Enough for Modeling Trust?

As for the growing trust literature in Information Technology and Artificial Intelligence (see Chapter 12 in this book and 'Trust: Theory and Technology group' site (http://www.istc.cnr.it/T3/), let us cite here only Marsh's thesis, which has been in fact the first attempt.

According to Marsh, *X* trusts *Y* if and only if '*X expects* that *Y* will behave according to *X*'s best interest, and will not attempt to harm *X*' (Marsh, 1994a; Marsh, 1994b).

This is a rather good definition of trust, but with some significant limits. First and foremost, it only refers to an attitude, the expectation of *X* that *Y* will behave according to *X*'s own interest. Therefore, in this definition, the notion of trust as decision and act, as reliance and counting on *Y*, is missed; and the idea of exposing oneself to failure or harm, of making oneself vulnerable by *Y* has also been overlooked.

Moreover, it is not clear whether *Y*'s behavior is (expected to be) intentional or not necessarily intentional. The first part of the definition is ambiguous about this: '*Y* will behave according . . .' – does this mean intentionally or accidentally? Will *Y* intentionally help *X*, or will his action just be factually exploited by *X*? The second sentence instead seems to refer explicitly to an intentional act, since it uses the term 'attempt'; thus the possible harm that *X* does not expect from *Y* would be intended, not accidental. However, Marsh's definition leaves open the possibility that two independent conditions for *X* to trust *Y* are there: on the one hand, *X* must expect *Y* to behave in a certain way, either intentionally or unintentionally; in addition, *X* must expect *Y* not to attempt (hence intentionally, by definition) to bring *X* any harm. In this interpretation, the second sentence of Marsh's definition would refer to something akin to the feeling of safety that is frequently linked with the notion of trust. However, it remains manifest that the exceeding ambiguity of this definition of trust in turn impinges on its applicability.

Finally, it is worth noting that the notion of 'interest' is not defined by Marsh,[9] so we can only assume it to be somehow close to other notions like welfare and goals.

1.5.5 Yamagishi: Mixing up the Act of Trusting and the Act of Cooperating

Following Yamagishi's interpretation of his comparative results between American and Japanese culture (Yamagishi and Yamagishi, 1994) (Yamagishi, 2003), what characterizes the latter is *assurance* rather than proper trust.

This means that the Japanese are more trusting (let us mark this use of the notion of trust as ω), i.e. more disposed to rely on the others (*X*'s side), than the Americans, when and if they feel protected by *institutional mechanisms* (authorities and sanctions).

Moreover, according to Yamagishi, the Japanese would tend to trust (in a different sense, that we shall indicate with λ) only when it is better for them to do so, because of the institutional or social costs associated with being 'untrusting' (in Yamagishi's own words), i.e. only for avoiding sanctions (*Y*'s side).

First of all, notice that here there is a misleading use of the term 'trust': in the first case (ω), it means that *X trusts in Y to do* α, i.e. *X* believes that *Y* is trustworthy and relies on *Y*; in the second use (λ), to trust is an act of cooperation, the proper contribution the agent should make to a well-codified social interaction.

These two uses must be distinguished. Obviously they are related, since in Japan *X* contributes/cooperates since he worries about institutional sanctions, and he *trusts* the others because he ascribes to them the same cultural sensibility and worries. Nonetheless, the two perspectives are very different: expectations about the others' behavior, on the one hand, and

[9] For a definition about the notion of *interest* as different from *goal* (and on trust relative to *interest protection*) see Chapter 2.

my own behavior of contributing, on the other, must be distinguished. We cannot simply label them both 'trust'.

Second, the confusion between 'tendency to trust' and 'tendency to cooperate/contribute', and between 'not trusting' and 'not cooperating/contributing' is misleading per se. If *X* cooperates just in order to avoid possible sanctions from the authority or group, trust is not involved. *X* does not contribute because he either trusts or does not trust the others, but rather for fear of sanctions – in fact, the only thing that *X* is trusting are the social authorities and their capacity for monitoring and sanctioning his conduct ((Castelfranchi and Falcone, 1998) (Falcone and Castelfranchi, 2001)). Calling this cognitive attitude 'tendency to trust' may be quite confusing. This is not simply a problem of terminology and conceptual confusion; it is a problem of *behavioral* notions that are proposed as *psychological* ones.[10]

Finally, here the concept of 'trusting' ends up losing its meaning completely. By missing the fundamental elements of having a positive evaluation of others and good expectations about their behavior, and because of these reasons relying on them and becoming vulnerable to them, this notion of trust comes to mean just to cooperate (in a game theoretical sense), to contribute to the collective welfare and to risk for whatever reason. The resulting equation 'Trust = to contribute/cooperate'; 'untrust = do not contribute/cooperate' is wrong in both senses: *there are cooperative behaviors without any trust in the others, as well as there being trust in the others in non-cooperative situations*.

We have to say that simply cooperating for whatever reason is not to trust. The idea that this *behavior* necessarily denotes trust by the agent and is based on this, so that the behavior can be used as a synonym of the attitude, is wrong. For example, as we have already said, worrying about institutional sanctions from the authority has nothing to do with putting one's trust *in* the other. Confusion between these two attitudes is fostered, among other things, by the fact that, usually, it is not specified in whom and about what a given subject trusts another, and based on what expectations and evaluations one has on the other. One should be clear in distinguishing between *X trusting the others* (possibly because he believes that they worry about the social authority and its possible sanctions), and *X doing something pro-collectivity* just because he worries about sanctions, not because he trusts the others.

Furthermore, it is wrong to assume that trust coincides with cooperation in these kinds of social dilemmas, and it is misleading to propose that trust consists in betting on some reciprocation or symmetric behavior. Here Yamagishi is clearly influenced by economists and their mental framework (see Chapter 8). As we will explain, trust also operates in completely different social situations.

Trust is not the feeling and disposition of the 'helper', but rather of the 'receiver' of some expected contribution to one's own actions and goals. Trust is the feeling of the helper only if the help (goal-adoption) is *instrumental* to triggering some action by the other (for example, some reciprocation). In this case, *X* is cooperating towards *Y* and trusting *Y*, but only because he is expecting something from *Y*. More precisely, the claim of interest for the economists is that *X is 'cooperating' because he is trusting* (in view of some reciprocation); he wouldn't cooperate without such a trust in *Y*.[11]

[10] Calling this behavior "trust behavior" is rather problematic for other reasons: it can be a behavior just relying in fact on the others' concurrent behavior, but unconsciously, without any awareness of 'cooperation'; as is the case – for a large majority of people – with paying taxes.

[11] In other words here we have a *double* and *symmetric* structure (at least in *X* mind) of goal-adoption and reliance (see later).

However, this is just a very peculiar case, and it is mistaken to take it as prototypical for founding the notion and the theory of 'trust' and of 'cooperation'. In general, a symmetric and reciprocal view of trust is unprincipled. In contrast, we can (and should) distinguish between *two basic constituents and moves of pro-social relations*:

- On the one side, *goal-adoption*, i.e. the disposition (and eventually the decision) of doing something *for* the other, in order to favor him.
- On the other side, *delegation*, i.e. the disposition (and eventually the decision) to count on the other, to delegate to him the realization of our goals and welfare (Castelfranchi and Falcone, 1998).

It is important to realize that this basic pro-social structure, which constitutes the nucleus of cooperation, exchange, and many other social relations, is *bilateral but not symmetrical.* In other words, pro-social bilateral relations do not start with reciprocation (which would entail some symmetry), nor with any form of exchange. The basic structure is instead composed by a social disposition and act of counting on the other and being dependent on him, thus expecting adoption from him (i.e. *trust*); this pro-social attitude will hopefully be matched by a disposition and an act of doing something for the other, of goal-adoption (see on this point Spinoza's notion of *benevolence*).[12]

Notice that the anti-social corresponding bilateral structure is composed of *hostility*, i.e. the disposition not to help or to harm, paired with *distrust* and *diffidence* from the other actor.

As this analysis should have made clear, benevolence and trust are not at all the same move or disposition (although both are pro-social and often combine together); they belong to and characterize two different although complementary actors and roles. Benevolence and trust are *complementary* and closely related, but they are also in part independent: they can occur without each other and can just be 'unilateral'. *X* can rely on *Y*, and trust him, without *Y* being benevolent towards *X*. Not only in the sense that *X*'s expectation is wrong and he will be disappointed by *Y*; but in the sense that *X* can successfully rely on *Y* and exploit *Y*'s 'help' without any awareness or adoption by *Y*. On the other hand, *Y* can unilaterally adopt *X*'s goals without any expectation from *X*, and even any awareness of such a help.

Moreover, both *trust and benevolence do not necessarily involve symmetric relations* between agents. It is possible to have cases of asymmetric trust[13] where only *X* trusts *Y*, while *Y* does not trust *X* (although he knows that *X* trusts him and *X* knows that *Y* does not trust her). And this holds both for trust about a specific kind of service or performance, as well as for generalized trust.

In addition, *trust does not presuppose any equality among the agents*, since there can be asymmetric power relationships between the trustor and the trustee: *X* can have much more power over *Y*, than *Y* over *X* (like it happens between parents and children). Analogously, goal-adoption can be fully asymmetrical, whenever *X* does something for *Y*, but not vice versa.

When there is a bilateral, symmetrical, and possibly reciprocal goal-adoption (i.e. the contribution of *X* to *Y*'s interests is also due to the help of *Y* towards the realization of *X*'s goals,

[12] On the contrary, 'justice' either is the rule of providing adoption or (if interpreted as 'fairness') it is also the rule of exchange, and thus it presupposes some reciprocity.

[13] This is for example in contrast with May Tuomela's account of Trust (Tuomela, 2003).

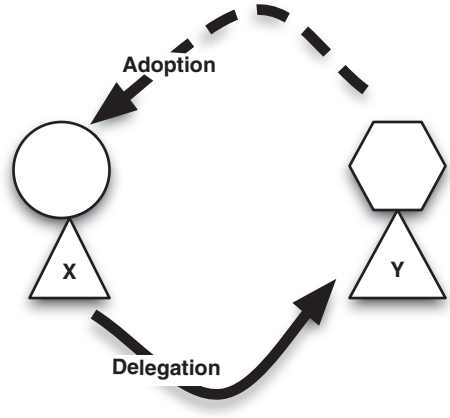

Figure 1.3 *X*'s Delegation meets *Y*'s Adoption

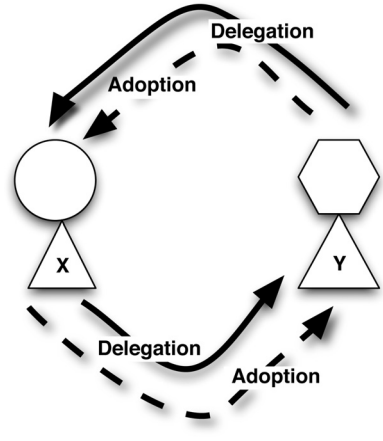

Figure 1.4 *X*'s Delegation meets *Y*'s Adoption, and vice versa: reciprocal trust

and vice versa), the structure presented in Figure 1.3 is doubled, as indicated in Figure 1.4. In this case, there is in fact trust/reliance from both sides and adoption from both sides.

Finally, it is worth noticing that even in *asynchronous* exchanges, when *X* acts *before Y* and *Y* acts only after *X*'s 'help', *Y* is trusting *X*. Not necessarily at the very moment of doing his own share, but before, when *Y* decides to accept *X*'s help and to rely on it.[14] Of course, in asynchronous 'exchanges' where *X* moves first, *X*'s trust in *Y* is broader and more risky: *X* has additionally to believe (before obtaining concrete evidence) that *Y* will do the expected action, whereas *Y* already has some evidence of *X*'s contribution (although this might be deceptive).

[14] For instance, *Y* has to believe that *X*'s help is good, is as needed, is convenient, is stable enough (it will not be taken back by *X*), and so on.

1.5.6 Trust as Based on Reciprocity

There is another important and recent family of definitions of trust, in which *trust denotes a behavior based on an expectation* (a view that we share), but both the behavior and the expectation are defined in a very restricted way. Let us take as the prototype of this approach the definition of trust provided by Elinor Omstrom and James Walker in their book *Trust and Reciprocity* (Omstrom and Walker, 2003): they define trust as 'the willingness to take some risk in relation to other individuals on the expectation that the others will reciprocate' (p. 382). This view in fact restricts the act of trusting to the act of 'cooperating', contributing, sustaining some cost in view of a future advantage that depends, in a strategic framework, also on the other's conduct; and it restricts the expectation to the expectation of reciprocation. By doing so, they exclude from the very notion and from the relations of trust all those cases that are not based at all on some exchange or cooperation; where *X* just counts on *Y*'s adoption of her goals, even in an asymmetric relationship, like in a son-mother relation, or in a request for help. We will extensively discuss these views in Chapter 8 about the notion of trust in economics and game theory.

1.5.7 Hardin: Trust as Encapsulated Interest

In Russell Hardin's view (Hardin, 2002), in part based on Baier's theory (Baier, 1986)):

> 'I trust you because I think it is in your interest to take my interests in the relevant matter seriously'.

This is the view of trust as *encapsulated interest*. I believe that you will take care of my interests, but I also believe that you will be rational, i.e. you will just follow your own interests; on such a basis I predict your favorable behavior: 'Any expectations I have are grounded in an understanding (perhaps mistaken) of your interests specifically with respect to me'. Expectations are not enough: 'The expectation must be grounded in the trustee's *concern* with the trustor's interests'[15].

With this very interesting notion of 'embedded interests', Hardin arrives close to capturing the crucial phenomenon of *social goal-adoption*, which really is foundational for any form of pro-sociality: from exchange to cooperation, from altruism to taking care of, and so on (Castelfranchi, 1989), (Conte and Castelfranchi, 1995), (Castelfranchi, 1998). In our vocabulary the fact 'that you encapsulate my interests in your own interests' means that you adopt my goals (for some reason of yours). Social goal-adoption is precisely the idea that another agent takes into account in his mind my goals (needs, desires, interests, projects, etc.), in order to satisfy them; he 'adopts' them as goals of himself, since he is an autonomous agent, i.e. self-driven and self-motivated (which does *not* necessarily mean being 'selfish'!), and he is not a hetero-directed agent, so that he can only act in view of and be driven by some internal purposive representation. Therefore, if such an internally represented (i.e. adopted) goal will be preferred to others, then he will happen to be self-regulated by my goal; for some motive of his own, he will act in order to realize my goal.

[15] See note 14.

As we shall discuss in a while, we agree with Hardin that there is a restricted notion of social trust which is based on *the expectation of adoption*, not just on the prediction of a favorable behavior of *Y*. When *X* trusts *Y* in a strict social sense and counts on him, she expects that *Y* will *adopt* her goal and that this goal will prevail, in case of conflict with other active goals of his. That is, *X* not only expects an *adoptive goal* by *Y*, but also an *adoptive decision and intention*. A simple regularity-based prediction or an expectation grounded on some social role or normative behavior of *Y* are not enough (and we agree with Hardin on this) for characterizing what he calls 'trust in a strong sense', which is the central nature of trust (Hardin, 2002), what we call *genuine social trust*.

However, to our mind, Hardin still fails to conceive a broad theory of goal-adoption, so that his notion of 'encapsulated interests' provides us with only a restricted and reductive view of it. Indeed, the theory of adoption is crucial, although not yet well understood, in sociology, economics and game theory (Tummolini, 2006), and in cooperation theory (Castelfranchi, 1997), (Tuomela, 1988, 1993). The fundamental point is to realize that *there are different kinds of goal-adoption depending on Y's motives (higher goals) for doing something 'for X'*, for spending resources in order to realize another agent's goal. Let us list these different cases:

1. Adoption can be just *instrumental* to *Y*'s personal and non-social goals; completely *selfish*. Like when we satisfy the need of chickens for food, only in order to later kill and eat them to our best satisfaction; or like when we enter a *do ut des* relation, like an economic exchange in Adam Smith's view (Smith, 1776).
2. Adoption can be *cooperative* in a strict sense. *X* and *Y* are *reciprocally dependent on each other* but just for one and the same goal, which constitutes their common goal. This is very different from other social situations, e.g. exchange, where there are two different and private/personal goals. Here instead the agents care for the same result in the world, and they need each other for achieving it. For this reason (being part of a necessarily common plan), each of them favors the actions and goals of the other within that plan, since he needs them and relies on them. In a sense, this adoption may be considered a sub-case of *instrumental* adoption, but it is definitely better to clearly distinguish (2) from (1). In fact, in (1) a rational agent should try to cheat, to avoid his contribution to the other: especially after having received *Y*'s service or commodity, and assuming no iteration of the interaction, *X* has no reason for doing her share, for giving *Y* what *Y* expects. On the contrary, in strict cooperative situations, based on real reciprocal dependence for the same objective, to cheat is self-defeating; without doing her share, *X* will not achieve her own (and common) goal (Conte and Castelfranchi, 1995), (Castelfranchi, Cesta, Conte, Miceli, 1993).
3. Finally, there is also non-instrumental, *terminal*, or altruistic adoption. The good of *X*, the realization of *X*'s needs, desires, and interests is an end *per se*, i.e. it does not need to be motivated by higher personal goals.

It is then just an empirical matter whether a behavior and cognitive structure as postulated in (3) does really exist in humans, or if humans are always motivated by selfish motives (although reduced to internal hidden rewards, like self-approval and esteem, or avoiding feelings of guilt

or pity), so that only 'pseudo-altruism' actually exists (Batson, 1991), (Lorini, 2005).[16] What really matters here is the well-defined *theoretical* possibility of having such an altruistic mind.

On this point, unfortunately Hardin seems to remain within the cultural framework of economics and game theory. It seems that his notion of 'self-interest' and 'rationality' necessarily collapse in 'selfishness'. In fact, missing a clear and explicit theory of purposive (internally goal-driven) behavior and thus of autonomous purposive agents, and an explicit theory of the goals of the agents (instead of just considering their quantitative dimension, i.e. utility), these approaches are not able to disentangle the fact that by definition such agents act for *their own* (internally represented and regulating) goals, choosing whatever options happen to maximize the achievement of those goals (that is, 'rationally'), from the wrong conclusion that *thus* they can only be 'selfish'. This is completely mistaken, since selfishness does not refer to autonomy, to non-exogenous but endogenous regulation, to being regulated by 'my own' goals; it refers instead to *a specific sub-family of the agent's goals*, which are non-adoptive, i.e. they do not aim to realize the goal of another agent.[17]

For example, one might build a robot, fully autonomous and rational in choosing among its goals, self-regulated by *its own* internal goals that, however, might just consist in the goal of helping its user, realizing all the desires or requests of the other, independently of any cost (including 'life') or 'personal' risks (which *subjectively* would not in fact exist at all, not having 'personal' goals, like safety or saving resources).

An important corollary of this view is that rationality has nothing to do with the nature or content of my goals; it cannot prescribe me my motives. A suicide or an altruist can take a perfectly subjectively rational decision, given their *unobjectionable* preferences. *Rationality is only instrumental: it is all about how we order and choose among our (conflicting) goals*.

In Hardin's perspective it seems that, whenever *Y* adopts *X*'s goals and does something *for X*, so that, being autonomous, he is adopting *X*'s goal for some reason, i.e. some of his *own* goals (Conte and Castelfranchi, 1995; Castelfranchi, 1998) this necessarily means that *X*'s goals are subordinated to *Y*'s personal and selfish convenience – an untenable claim, as discussed above.

[16] However, we do not believe it to be so. In our view, psychology still fails to model a very fundamental and obvious distinction: the fact that I have a positive/likeable expectation while deciding to do a given action does *not* entail that I am doing that action *for* that outcome, in order to achieve it, being motivated by it. As it happens I am not motivated by bad *expected* results, it happens also that I am not necessarily motivated by good *expected* results. It is true that my motivating goal (the goal in view of which I act, and which is both 'necessary' and 'sufficient' for intending to act; (Lorini, Marzo, Castelfranchi, 2005) must be among the positive (goal-satisfying) expected results, but it is not true that all the positive expected results motivate me, being necessary or sufficient for deciding to perform that action. Seneca had clarified this point already 2000 years ago. It would be nice if psychology would at last acknowledge and model such a distinction. More importantly, Seneca's claim effectively defeats the 'pseudo-altruism' thesis and opens the way – at least in principle – to true psychological altruism. "*Sed tu quoque' inquit 'uirtutem non ob aliud colis quam quia aliquam ex illa speras uoluptatem. Primum non, si uoluptatem praestatura uirtus est, ideo propter hanc petitur; non enim hanc praestat, sed et hanc, nec huic laborat, sed labor eius, quamuis aliud petat, hoc quoque adsequetur*" (*De vita beata*, IX). In a few words: Do we cultivate our virtues just because we wish to obtain some pleasure? The fact that virtue gives some pleasure does not mean that we follow it *because of* this. The pleasure is just an additional result, not our aim or motive. We will get it while pursuing another aim, that is, virtue.

[17] Otherwise, *any* act would be 'selfish', egoistic. But this is *not* our usual notion of selfishness at all, and it does not correspond to our intuitions about our own and each other motives. Rather, this is the expression of a restricted philosophical view, embedded in a cynical understanding of human nature.

In Hardin there is also a strong but quite arbitrary restriction (although typical of the game-theoretic framework[18]) on *Y*'s *motives* for caring about *X*'s interests and doing something for *X*, to *Y*'s desire to maintain that relationship: 'I trust you because I think it is your interest to attend to my interests in the relevant matter. This is not merely to say that you and I have the *same* interests. Rather, it is to say that you have an interest in attending to *my* interest because, *typically, you want our relationship to continue*' (Hardin, 2005, Ch.1).

Here we agree with Hardin on the claim that trust is a three-part relation: *X trusts Y to do* α (this relates also to the analysis of Baier (Baier, 1986), and Luhmann (Luhmann, 1979)). In our own terminology, this would translate as: *X trusts Y as for something*, e.g. doing α (we call α the *delegated task* (Castelfranchi and Falcone, 1998)). Many authors do not realize the importance of the third element (indeed, even the context deserves to be modeled explicitly: *X trusts Y in context C to do* α) and this creates several problems; for example, they cannot develop a theory of disomogeneous trust evaluation (*X* trusts *Y* for one task but not for another one), or a theory of how we transfer trust in *Y* for one performance to trust in *Y* for another performance – an issue of paramount importance in marketing (see Chapter 8). Thus we agree with Hardin that there is no two-part trust or one-part trust; they are just elliptical instances of a more complex relation.

However, we disagree about the rigid use that Hardin makes of this triadic view. He seems to reject any theory of trust generalization, transfer, and abstraction, and focus instead only on a specific personal relation (i.e. trust in a *specific* person, not for example in a 'category'), and only for a specific action α (not for other similar actions, or for a class of actions). This is probably due to his quest for an understanding of *Y*'s specific incentives and interests in attending *X*'s interest. So, although Hardin admits the importance and the advantages of a 'general atmosphere of trustworthiness', in his model this would necessarily and only be a rich network of specific trust relationships among particular individuals: lots of people, each trusting other 'particular' people (Hardin, 2005 (p. 179)). But this, of course, is not what it is usually meant by 'generalized trust'.

In contrast, our approach will be shown adequate to account for trust transmission, generalization, and vagueness (for any value of *Y*; for any value of *X*; for any value of α; see Chapter 6 for details). We deny that the theory of generalized or diffuse trust (the so called 'social capital') is another and different model from the theory of interpersonal trust. In our perspective, the former must be built upon and derived from the latter. On the one side, trust capital is a macro, emerging phenomenon; but it must be understood also in terms of its micro foundations. On the other side, conceptually speaking, the notion of trust is just one and the same notion, at different levels of abstraction and generalization.

Another significant disagreement that we have with Hardin is that he decided to consider as 'trust' only the epistemic/doxastic mental attitude of the trustor, i.e. his beliefs. Hardin's book begins with this statement: 'To say that I trust you (. . .) simply means that I *think* you will be trustworthy toward me'; 'To say that I trust you (. . .) is to say nothing more than that I *know or believe certain things about you*' (Hardin, 2005).

Moreover, he claims that the declarations *I believe you are trustworthy* and *I trust you* are equivalent. This is clearly false. The first sentence is the paraphrase of only one of at least two

[18] Hardin cites Thomas Schelling (1960, 1980, 134-5): "Trust is often achieved simply by the continuity of the relation between parties and the recognition by each that what he might gain by cheating in a given instance is outweighed by the value of the tradition of trust that makes possible a long sequence of future agreement".

different readings of the second sentence. *I trust you* can mean *I believe you are trustworthy*, but it can also mean much more than this: in particular, it can mean *I have decided to rely on you, to depend and count on you, since I believe you to be trustworthy.*[19] Trust in our model is not only a doxastic mental attitude, but also a *decision* and *intention*, and the *subsequent action* based upon these mental dispositions. There is no contradiction between the so-called 'cognitive' and 'behavioral' notion (acting on trust). In contrast, for Hardin, trust 'if it is cognitive is not behavioral' (Hardin, 2005).[20] In our model the two notions are embedded one into the other; and the explicit theory of the relations between these layers and of the transition from one to the other is important.

Based on this view is the idea that in principle one cannot really *decide* to trust somebody else; trust 'is not a matter of choice', insofar as we refer to trust as a doxastic attitude. This is because, as a matter of fact, one cannot decide to believe something[21]; but this fact is not true for the second layer of trust, concerning the decision to delegate and the act based upon this decision. While trusting you I can really 'decide' to take some risk (which I mentally consider) and to make myself vulnerable to you. So much so that, indeed, I can later come to regret my decision, and blame myself for that choice.

It follows that trust can be (and usually is) rational on both these two levels: as a belief, and as a decision and action. *Epistemic rationality* consists in forming a justified belief on the basis of good evidence and reliable sources. *Decision rationality* consists in taking a decision on the basis of a correct calculation of values (outcomes) and of their (limited) optimization. In our model, trust can be rational at both levels: grounded as a belief (evaluation and prediction), and optimal as a choice.

The last important difference between Hardin's analysis and our model is that Hardin completely puts aside the issue of 'competence', overlooking all those cases in which we put our trust and reliance on the trustee's ability, expertise, quality of service, and so on. In his own words: 'I will usually assume through this book that competence is not at issue in the trust relationships under discussion' (Hardin, 2005, Introduction). Unfortunately, this assumption is untenable, since the competence aspect cannot be either marginalized or rigidly separated from trust in *Y*'s reliability (see Section 2.2.5 for more details).

1.5.8 Rousseau: What Kind of Intention is 'Trust'?

A very good (indeed, our favorite) definition of trust, based on a large interdisciplinary literature and on the identification of fundamental and convergent elements, is the following: '[Trust is] a psychological state of a trustor comprising the intention to accept vulnerability in a situation involving risk, based on positive expectations of the intentions or behavior of the trustee' (Rousseau *et al*., 1998).

---

[19] This is the meaning of 'trust' in expressions like: "While trusting *Y* you expose yourself, you risk very much!"; "My poor friend, how could you trust him?!"; "Trust me! You can trust me!" "OK, I trust you!".

[20] This claim can be correct if 'behavioral' means a behavioristic definition (like in Yamagishi's approach); but it is wrong if it is aimed at excluding trust as decision and action. Actually 'Action' is a cognitive notion.

[21] Nonetheless, there may be trust attitudes based on 'acceptances' and not on 'beliefs'; and 'acceptances' can be intentionally assumed. In this perspective, I can also 'decide' to trust you, in the sense that I decide to presume that you are reliable and competent, and to act on the basis of such an assumption, i.e. "as if" I believed you to be reliable and competent (Cohen, 1992) (Engel, 1998).

What we like here is the idea of a composite psychological state (although it is not fully characterized), which does not only include 'positive expectations', but where these expectations are the *base* for the intention to make oneself vulnerable. This crucial link is made explicit in this definition, so it is less important that other beliefs are ignored (like the fact that trust is also an appraisal), that the 'competence' and 'ability' of *Y* remain implicit, or that there seems to be no trust before and without intention. Notice that – given this definition – since trust necessarily includes the 'intention', the idea of trust might not be enough to entrust *Y*, would be contradictory. One could never say: 'I trust *Y* but not enough', or 'I trust *Y* but I do not intend to rely on him'.

This brief survey of various definitions of trust from different disciplines was just for highlighting the main limits of the current (mis)understanding of this notion, but also a lot of very important intuitions that – although partial and even contradictory – deserve to be preserved, well defined, and coherently organized.

On the basis of these results, it is now time to start introducing in a more explicit and systematic way our layered model of trust, which is what we shall do in the next chapter.[22] We will encounter and discuss other definitions and models throughout the book.

References

Baier, A. Trust and antitrust, *Ethics*, 96 (2): 231–260, 1986.

Batson, C.D. (1991) *The Altruism Question: Towards a social social-psychological answer*, Hillsdale, NJ: Erlbaum.

Busacca, B. and Castaldo, S. Trust in market relationships: an interpretative model, *Sinergie*, 20 (58): 191–227, 2002..

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

Castelfranchi, C. Paradisi artificiali: prolegomeni alla simulazione della interazione sociale. *Sistemi Intelligenti*, I (1): 123–165, 1989.

Castelfranchi, C. (1997) Principles of (individual) social action. In Tuomela, R. & Hintikka, G. (eds) *Contemporary Action Theory*. Kluwer.

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

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

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., Cesta, A., Conte, R., Miceli, M. Foundations for interaction: the dependency theory, *Advances in Artificial Intelligence*. Lecture Notes in Computer Science; Vol. 728: 59–64, 1993. ISBN:3-540-57292-9.

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

Castelfranchi, C. and Falcone, R. Founding autonomy: the dialectics between (social) environment and agent's architecture and powers. *Agents and Computational Autonomy*, 40–54, 2003.

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.

[22] Later, (see Chapter 8), we come back to the scientific literature, in order to dedicate special attention to economic and game-theoretic accounts of trust, considering both recent developments in this domain, and the old Deutsch's definition of this notion.

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.

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

Conte, R. and Castelfranchi, C. (1995) Minds is not enough. Pre-cognitive bases of social interaction. In: Gilbert, N., Doran, J., *Simulating Societies*, UCL, London.

Cummings, L. L. and Bromiley, P. (1996) The organizational trust inventory (OTI): development and validation. In R. Kramer, and T. Tyler (eds.), *Trust in Organizations* (pp. 302–330). Thousand Oaks, CA: Sage.

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.

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., 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. and Castelfranchi, C. (2001) The human in the loop of a delegated agent: The theory of adjustable social autonomy, *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, Special Issue on 'Socially Intelligent Agents - the Human in the Loop, 31 (5): 406–418, September 2001.

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

Falcone, R. and Castelfranchi, C. (2001) 'The socio-cognitive dynamics of trust: does trust create trust?' In R. Falcone, M. Singh, Y.H. Tan, eds., *Trust in Cyber-societies. Integrating the Human and Artificial Perspectives*, Heidelberg, Springer, LNAI 2246, 2001, pp. 55–72.

Falcone, R. and Castelfranchi, C. (2003) A belief-based model of trust, in Maija-Leena Huotari and Mirja Iivonen (eds) *Trust in Knowledge Management and Systems in Organizations*, Idea Group Publishing, pp. 306–343.

Gambetta, D. (1988) 'Can we trust trust?' In *Trust: Making and Breaking Cooperative Relations*, edited by D. Gambetta. Oxford, Blackwell.

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

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

Johannisson, B. (2001) Trust between organizations: state of the art and challenges for future research, paper presented at the EIASM Workshop on Trust Within and Between Organizations, Vrije Universiteit Amsterdam.

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

Jones, K. Trust as an affective attitude, *Ethics* 107: 4–25, 1996.

Lorini, E., Marzo, F., Castelfranchi, C. (2005) A cognitive model of the altruistic mind. Boicho Kokinov (eds.), *Advances in Cognitive Economics*, NBU Press, Sofia, Bulgaria, pp. 282–294.

Luhmann, N. (1979) *Trust and Power*, John Wiley & Sons Ltd, New York.

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

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.

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

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

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

Moorman, C., Zaltman, G., and Deshpandé, R. (1992) Relationships between providers and users of market research: the dynamics of trust within and between organizations. *Journal of Marketing Research* 29 (3) : 314–328, August, 1992.

Mutti, A. (1987) La fiducia. Un concetto fragile, una solida realta, in Rassegna italiana di sociologia, pp. 223–247.

Ostrom E. and Walker J. (eds.) (2003) *Trust and Reciprocity: Interdisciplinary Lessons from Experimental Research*, Russell Sage Foundation, New York, NY, 409 pp., ISBN 0-87154-647-7.

Rousseau, D. M., Burt, R. S., and Camerer, C. Not so different after all: a cross-discipline view of trust. *Journal of Academy Management Review*, 23 (3): 393–404, 1998.

Seneca, *De vita beata*, IX.

Schelling, T. (1960, 1980) *The Strategy of Conflict*, Harvard University Press.

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

Tummolini, L. and Castelfranchi, C. (2006) The cognitive and behavioral mediation of institutions: Towards an account of institutional actions. *Cognitive Systems Research*, 7 (2–3).

Tuomela, R. and Miller, K. We-Intentions, *Philosophical Studies*, 53: 115–137, 1988.

Tuomela, R. What is cooperation. *Erkenntnis*, 38: 87–101, 1993.

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

Williamson, O. E. (1993) Calculativeness, trust, and economic organization, *Journal of Law and Economics*, XXXVI: 453–486, April 1993.

Yamagishi, T. and Yamagishi, M. (1994) Trust and commitment in the United States and Japan. *Motivation and Emotion*. 18, 129–166.

Yamagishi, T. (2003) Cross-societal experimentation on trust: A comparison of the United States and Japan. In E. Omstrom and J. Walker, eds., *Trust, Reciprocity: Interdisciplinary Lessons from Experimental Research* NY, Sage, pp. 352–370.