

Social Trust: A Cognitive Approach

Cristiano Castelfranchi and Rino Falcone

National Research Council - Institute of Psychology *

Unit of "AI, Cognitive Modelling and Interaction"

Roma - Italy

Abstract

After arguing about the crucial importance of trust for Agents and MAS, we provide a definition of trust both as a mental state and as a social attitude and relation. We present the mental ingredients of trust: its specific beliefs and goals, with special attention to evaluations and expectations. We show the relation between trust and the mental background of delegation. We explain why trust is a bet, and implies some risks, and analyse the more complex forms of social trust, based on a theory of mind and in particular on morality, reputation and disposition, and authority (three party trust). We explain why promises, contracts, authorities can increase our trust by modifying our mental representations. We present a principled quantification of trust, based on its cognitive ingredients, and use this "degree of trust" as the basis for a rational decision to delegate or not to another agent. We explain when trust is rational, and why it is not an irrational decision by definition. We also criticise the economic and game-theoretic view of trust for underestimating the importance of cognitive ingredients of trust and for reducing it to subjective probability and risk. The paper is intended to contribute both to the conceptual analysis and to the practical use of trust in social theory and MAS.

1. Premise: the importance of trust

As it has been written in the call of the original workshop "In recent research on electronic commerce trust has been recognized as one of the key factors for successful electronic commerce adoption. In electronic commerce problems of trust are magnified, because agents reach out far beyond their familiar trade environments. Also it is far from obvious whether existing paper-based techniques for fraud detection and prevention are adequate to establish trust in an electronic network environment where you usually never meet your trade partner face to face, and where messages can be read or copied a million times without leaving any trace. (...) With the growing impact of electronic commerce distance trust building becomes more and more important, and better models of trust and deception are needed. One trend is that in electronic communication channels extra agents, the so-called Trusted Third Parties, are introduced in an agent community that take care of trust building among the other agents in the network. But in fact *different kind of trust are needed* and should be modelled and supported:

- trust in the environment and in the infrastructure (the socio-technical system);
- trust in your agent and in mediating agents;
- trust in the potential partners;
- trust in the warrantors and authorities (if any).

The notion of trust is also important in other domains of agents' theory, beyond that of electronic commerce. It seems even foundational for the notion of "agency" and for its defining relation of acting "on behalf of" [Cas2]. For example, trust is relevant in Human-Computer interaction, e.g., the trust relation between the user and her/his personal assistant (and, in general, her/his computer). It is also critical for modelling and supporting groups and teams, organisations, co-ordination, negotiation, with the related trade-off between local/individual utility and global/collective interest; or in modelling distributed knowledge and its circulation. In sum, the notion of trust is crucial for all the major topics of Multi-Agent systems. What is needed is a general and principled theory of trust, of its cognitive and affective components, and of its social functions. A theory has to be developed to answer questions like the following: When is trust rational? When is it over-confidence and risky? When is trust too weak and when do we waste time on redundant control

* This research has been developed within the agreement between CNR and Provincia Autonoma di Trento, research project on "Applicazioni avanzate di informatica".

We would like to thank Maria Miceli, YaoHua Tan, Maj Bonniver Tuomela, Giovan Francesco Lanzara, and Walter Thoen for useful comments and discussions.

mechanisms or loose good opportunities by not taking advantage of sufficient trust levels?” [CfP].

In this contribution we attempt to answer these questions and provide a basic theory of trust. We will put aside trust in object, events and tools [Cas6] -although we provide a very general characterisation of trust- and we do not discuss the affective components of it [Tha]. We present a *cognitive model of trust* in term of necessary mental ingredients (beliefs and goals) and decision to delegate.

We stress the importance of this explicit cognitive account for trust in three ways. First, we criticize the game-theoretic view of trust which is prisoner of the Prisoner Dilemma mental frame and reduces trust simply to a probability or perceived risk in decisions [Wil]. Second, we found the quantitative aspects of trust (its strength or degree) on those mental ingredients (beliefs and goals) and on their strength. Third, we claim that this cognitive analysis of trust is fundamental for distinguishing between internal and external attribution which predict very different strategies for building or increasing trust; for founding mechanisms of image, reputation, persuasion, argumentation in trust building. Finally, we show that this cognitive anatomy is important for deeply understanding the relationship between trust, delegation, and its different levels and kinds.

Delegation is in fact a multi-agent oriented act and relation (usually a "social" act and relation), which is based on a *specific set of beliefs and goals and on a decision*. This complex and typical mental state is "trust". In order to delegate a task to some agents y (collaborator) I have to *believe* that it is able to do what I need (competence), and that it will actually do that (predictability). Now, these beliefs, with their degree of uncertainty, represent precisely my trust in y , and my action of delegating an action α or a goal g to y , represents precisely my action of entrusting y for α/g .

Consider for example the basic definition of trust provided in the classic book of Gambetta and accepted the great majority of the authors [Gam]:

“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, and it stresses that trust is basically an estimation, an opinion, an evaluation, i.e. a belief. However, it is also quite a poor definition, since it just refers to one dimension of trust (predictability), while ignoring the “competence” dimension; it does not account for the meaning of “I trust B” where there is also the *decision* and the *act* of relying on B; and it doesn't explain what is such an evaluation made of and based on: in fact, the *subjective probability* melts together too many important parameters and beliefs, which are very relevant in social reasoning.

2. What trust is

In this section we will carefully analyse the ingredients necessary to have the mental state of trust, i.e. the components and sources of that estimated “subjective probability” and of that expectation about the profitable behaviour of another agent. We will specify which beliefs and which goals characterise x 's trust in another agent y . Given the overlap between trust and reliance/delegation, we need also to clarify their relationship.

2.1 A mental attitude

Assertion 1: *Only a cognitive agent can “trust” another agent: only an agent endowed with goals and beliefs.*

First, one trusts another only relatively to a goal, i.e. for something s/he want to achieve, that s/he desires. If I don't potentially have goals, I cannot really decide, nor care about something (welfare): I cannot subjectively “trust” somebody. Second, trust itself *consists of* beliefs.

Assertion 2: Trust basically is a *mental state*, a complex *attitude* of an agent x towards another agent y about the behaviour/action α relevant for the result (goal) g .

- x is the *relying agent*, who feels trust; it is a *cognitive agent* endowed with internal explicit goals and beliefs;
- y is the agent or entity which is trusted; y is not necessarily a cognitive agent. So
- x trusts y about g/α and for g/α ; x trusts also “that” g will be true.

Since y 's action is useful to x , and x is relying on it, this means that x is "delegating" some action/goal in her own plan to y . This is the strict relation between trust and reliance or delegation.

Assertion 3: *Trust is the mental counter-part of delegation.*

Given this strict relation and the foundational role of delegation (see 1.3) we need to define delegation and its levels (2.2); and to clarify also differences between delegation and trust (2.3). After this we will go back to examine the mental ingredients of trust, and their social significance.

2.2 What delegation is

Assertion 4: *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 on y . x plans to achieve g through y . So, she is constructing a multi-agent plan and y has a share in this plan: y 's delegated task is either a state-goal or an action-goal [Cas1].*

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.

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 her action. 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 to the bird an action in his plan; and the bird is unconsciously and unintentionally collaborating with the hunter's plan.

In stronger forms of delegation x can herself eliciting, inducing the desired y 's behaviour 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.

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.

2.3 Trust and delegation

Although we claim that trust is the mental counter-part of delegation, i.e. that it is a structured set of mental attitudes characterising the mind of a delegating agent/trustor, however obviously there are important differences, and some independence, between trust and delegation. Trust and delegation are not the same. The word "trust" is also ambiguous, it denotes both the simple evaluation of y before relying on it (we will call this "core trust"), the same plus the decision of relying on y (we will call this part of the complex mental state of trust "reliance"), and the *action* of trusting, depending upon y (this meaning really overlaps with "delegation" and we will not use the term "trust" for this).

Trust is first of all *a mental state*, an *attitude* towards an other agent (usually a social attitude). We will use the three argument predicate **-Trust (x y τ)**- to denote a specific *mental state* compound of other more elementary mental attitudes (beliefs, goals, etc.). While we use a predicate **Delegate (x y τ)** to denote the *action* and the resulting relation between x and y .

Delegation necessarily is an *action*, a result of a decision, and it also creates and is a (*social*) *relation* among x , y , and α . The external, observable action/behaviour of delegating either consists of the action of provoking the desired behaviour, of convincing and negotiating, of charging and empowering, or just consists of the action of doing nothing (omission) waiting for and exploiting the behaviour of the other. Indeed, will we use trust and reliance only to denote the mental state preparing and underlying delegation (*trust* will be both: the small nucleus and the whole)¹.

Assertion 5: *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).*

¹ 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.

So, trust is normally *necessary* for delegation, but it is not *sufficient*: delegation requires a richer decision.

Assertion 6: *There may be delegation without trust: these are exceptional cases in which either the delegating agent is not free (coercive delegation ²) or he has no information and no alternative to delegating, so that he must just make a trial (blind delegation).*

The decision to delegate has no degrees: either x delegates or x does not delegate³. 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.

2.4 Basic Beliefs in Trust

Let us start from the most elementary case of trust and delegation: it is useful in fact to identify what ingredients are really basic in the mental state of trust.

First, x has a **goal** g x tries to achieve by using y : this is what x would like to "delegate to" y , its "task" [Fal]. Then x has some specific beliefs:

1. "Competence" Belief: a *positive evaluation* of y is necessary, x should believe that y is useful for this goal of its, that y can produce/provide the expected result, that y can play such a role in x 's plan/action, that y has some function.

2. "Disposition" Belief: Moreover, x should think that y not only is able and can do that action/task, but y actually will do what x needs. With cognitive agents this will be a belief relative to their *willingness*: this make them predictable [Mic].

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

Esteem and Expectation

Let us stress the nature of the above beliefs about y . They are *evaluations* and *positive expectations*, not "neutral" beliefs. In trusting y , x believes that y has the right qualities, power, ability, competence, and disposition for g . Thus the trust that x has in y is (clearly and importantly) part of (and is based on) her/his esteem, her/his "image", her/his reputation [Das, Rau].

There is also a "positive expectation" about y 's power and performance.

Assertion 7: A *positive expectation* is the combination of a goal and of a belief about the future (prediction): *x both believes that g and x desires/intends that g .*

In this case: x both *believes* that y can and will do; and x *desires/wants* that y can and will do⁴.

2.5 Trust and Reliance

The kernel ingredients we just identified are not enough for arriving to a delegation or reliance disposition. At least a third belief is necessary for this:

3. Dependence Belief: x believes -to trust y and delegate to it- that either x needs it, x depends on it (*strong dependence* [Sic], or at least that it is better to x to rely than do not rely on it (*weak dependence*, [Jen]).⁵

² Suppose that you don't trust at all a drunk guy as a driver, but you are forced by his gun to let he drive your car.

³ In general, in (delegates x y), y is a set of agents (either in "and" or "or" relation). If x delegates to all agents in y the same task, in fact x realizes -considering each agent in y - a delegation with degrees. However for simplicity, we will consider just the case in which y is one alone agent.

⁴ The fact that when x trusts y , x has a positive expectation, explains why there is an important relationship between trust and hope, since also **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 behaviour).

⁵ Among the basic ingredients there is a 4th belief -that is especially important for the affective aspects of trust, but not only- we will not consider here: the Belief of Unharmfulness. When x trusts y , x feels/believes that it is not harmful, threatening me, dangerous or hostile (for ex. with people x believes that s/he is not an enemy, or has no reason (envy, hate, etc.) to hurt x : diffidence enhances diffidence, sympathy elicits sympathy, trust induces trust in the other. There is stronger form of this in social trust, which is the feeling about the good disposition of the other,

In other terms, when x trusts on someone, x is in *a strategic situation* [Deu2]: x believes that there is interference [Cas7] and that her rewards, the results of her projects, depend on the actions of another agent y .

These beliefs (plus the goal g) define her "trusting y " or her "**trust in y** "⁶ in delegation. However, another crucial belief arises in x 's mental state -supported and implied by the previous ones:

4. **Fulfilment Belief**: x believes that g will be achieved (thanks to y in this case)⁷. This is the "**trust that**" g .

Thus, *when x trusts y for g , it also has some trust that g* . When x decides to trust, x has also the new goal that y performs α , and x rely on y 's α in her plan (delegation) (about this decision see 6.). In other words, on the basis of those beliefs about y , x "leans against", "count on", "depends upon", "relies on", in other words x practically "trusts" y . Where -notice- "to trust" does not only means those basic beliefs (the core) but also the decision (the broad mental state) and the act of delegating.

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

This decision is the second crucial component of the mental state of trust: let us call this part "reliance", and the first part "core trust" (and the whole picture mental state of trust and the behaviour "delegation"). Using Meyer, van Linder, van der Hoek et al.'s logics [Mey, Lin], and introducing some "ad hoc" predicate (like WillDo) we can summarise and simplify the mental ingredients of trust as follows:

$G_0: \text{Goal}_X(g)$ $PE_1 \left[\begin{array}{l} B_1: B_X \text{Can}_Y(\cdot, g) \\ G_1: W_X \text{Can}_Y(\cdot, g) \end{array} \right. \quad \text{(Competence)}$ $PE_2 \left[\begin{array}{l} B_2: B_X \langle \text{WillDo}_Y(\cdot) \rangle g \\ G_2: W_X \langle \text{WillDo}_Y(\cdot) \rangle g \end{array} \right. \quad \text{(Disposition)}$	Core Trust
$B_3: B_X \text{Dependence}_{XY}(\cdot, g) \quad \text{(Dependence)}$ $G_3: \text{Goal}_X \neg(\langle \text{WillDo}_X(\cdot) \rangle g)$ $G_4: \text{Goal}_X \langle \text{WillDo}_Y(\cdot) \rangle g$	Reliance

where: PE means positive expectation, B is the believe operator (the classical doxastic operator), and W is the wish operator (a normal modal operator).

Wishes are the agents' desires, they model the things the agents like to be the case; the difference between wishes and goals consists in the fact that goals are selected wishes.

The fulfilment belief derives from the formulas in the above schema.

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

Risk, Investment and Bet

Any act of trusting and relying implies some bet and some risk [Luh]. In fact, x might eventually be disappointed, deceived and betrayed by y : her beliefs may be wrong. At the same time x bets

his benevolence, or intention to adopt our goals and interest, and this support the belief and feeling of unharfulness.

⁶ 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 better analyse this important distinction in 3.2 about social trust.

⁷ The trust that g does not necessarily requires the trust in y . I might ignore which are the causal factors producing or maintaining g true in the world, nevertheless I may desire, expect and trust that g happens or continue. The Trust that g , per se, is just a -more or less supported- subjectively certain positive expectation (belief conform to desire) about g .

something on y . First, x renounced to (search for) possible alternatives (for ex. other partners) and x might have lost her opportunity: thus x is risking on y the utility of her goal g (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 it at her disposal (for ex. x paid for y or for its service); now this investment is a real bet [Deu] on y . Thus, to be precise we can say that:

Assertion 8: *When x trusts y there are two risks: a) the risk of failure, the frustration of g (possibly for ever, and possibly of the entire plan containing g);⁸ b) the risk of wasting the efforts.* Not only x risks to miss g (*missed gains*) but x also risks to waste her investments (*loss*).

As for the first risk, it is important to notice that:

Assertion 9: *trusting and betting on y might increase x 's dependence on y .*

In fact, if initially x might have alternatives to y (rely on z or w) after its choice (and perhaps because of this choice) z and w might be no more at its disposal (for example they might be busy); this means that x 's alternatives means (partners) for g are reduced and then x 's dependence on y has increased [Sic].

The act of trusting/reliance is a real wager, a risky activity: it logically presupposes some uncertainty, but it also requires some **predictability** of y , and usually some degree of trust in y .

3. Social Trust: Trusting Cognitive Agents ⁹

When applied to cognitive, intentional agents, the basic beliefs of trust need to be articulated in and supported by other beliefs. In fact, how to predict/expect that an intentional agent will do something, unless on the basis of the 'intentional stance', i.e. on the basis of beliefs about its motives, preferences, intentions, and commitments?

All this must be combined with different kinds of delegation. There are various kinds, levels and specific ingredients of trust in relation to the kind of delegation and the degree of autonomy.

Let us in particular see *weak Vs strong delegation*, and something about *open delegation* and *delegation of control*. These two forms of delegation are based on specific varieties of trust, with specific mental ingredients.

3.1 Trust in Weak Delegation

Weak delegation does not presuppose any agreement, deal or promise: for example, x weakly delegate when, being at a bus stop, x relies on another person to raise his arm and stop the bus, predicting that he will do this, and risking to miss her bus.

When applied to a cognitive, intentional agent, weak delegation implies that the "will-do" belief be articulated in and supported by a couple of other beliefs (that will continue to be valid also in strong delegation):

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

6. Persistence Belief. x should also believe that y is stable enough in his intentions, that has no serious conflicts about α (otherwise he might change his mind), or that y is not unpredictable by character, etc.¹⁰

When x relies on y for his action, x is taking advantage of his independent goals and intentions, predicting his behaviour on such a basis, or x is herself inducing such goals in order to exploit his behaviour. In any case x not only believes that y is able to do and can do (opportunity), but also that y will do because he is committed to this intention or plan (not necessarily to x).

⁸ Moreover there might be not only the frustration of g , the missed gain, but there might be additional damages as effect of failure, negative side effects: the risks in case of failure are not the simple counterpart of gains in case of success.

⁹ In [Cas6] we analyzed also trust in tools and objects.

¹⁰ Beliefs 5. and 6. imply some beliefs about y 's motives: intention is due to this motive, and persistence is due to preferences between motives. However this Motivation belief will be more important later, in Adoption-based trust.

7. Self-confidence Belief. x should also believe that y knows that he can do α . Thus he is self-confident. It is difficult to trust someone that does not trust himself! ¹¹

Let's *simplify* and formalise this.

Introducing some "ad hoc" predicate (like WillDo, or Persist)¹² in the logics of [Mey, Lin], we might characterise **social trust mental state** as follows:

$$\text{Trust}(X,Y,\tau) = \text{Goal}_X \supset \text{B}_X \text{PracPoss}_Y(\alpha, g) \supset \text{B}_X \text{Prefer}_X(\text{Done}_Y(\alpha, g), \text{Done}_X(\alpha, g)) \supset (\text{B}_X(\text{Intend}_Y(\alpha, g) \supset \text{Persist}_Y(\alpha, g)) \supset (\text{Goal}_X(\text{Intend}_Y(\alpha, g) \supset \text{Persist}_Y(\alpha, g))))$$

Where: $\text{PracPoss}_Y(\alpha, g) = \langle \text{Do}_Y(\alpha) \rangle g \supset \text{Ability}_Y(\alpha)$. To formalize results and opportunities this formalism borrow constructs from dynamic logic: $\langle \text{Do}_i(\alpha) \rangle g$ denotes that agent i has the opportunity to perform the action α in such a way that g will result from this performance.

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

3.2 Internal attribution of trust (trustworthiness)

We should distinguish between trust ‘in’ someone or something that has to act and produce a given performance thanks to its *internal* characteristics, and the global trust in the global event or process and its result which is also affected by external factors like opportunities and interferences. Trust *in* y (for example, ‘social trust’ in strict sense) seems to consists in the two first prototypical beliefs/evaluations we identified as the basis for reliance: *ability/competence* (that with cognitive agents includes self-confidence), and *disposition* (that with cognitive agents is based on willingness, persistence, engagement, etc.). Evaluation about *opportunities* is not really an evaluation about y (at most the belief about its ability to recognize, exploit and create opportunities is part of our trust ‘in’ y). We should also add an evaluation about the probability and consistence of obstacles, adversities, and interferences.

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

We didn't completely model the important ingredients/component of this trust. For example very important aspects, and in some sense the top of trust in y , are the idea that

- “ y will do his best for achieving g ”, he is very concerned and engaged, and will put any effort;
- or the idea that y is reactive, alerted and smart enough for both perceive, exploit and create positive opportunities for success, and for detecting or predicting and cope with obstacles and negative interferences;
- and what we later call the *pro-social attitude* of y .

The distinction between internal Vs external attribution is important for several reasons.

- First, to better capture the meaning of trust in several common sense and social science uses.
- Second to understand the precise role of that nucleus of trust that we just mention in terms of “unharmfulness”, sense of safety, perception of goodwill (see later).
- Third, to better understand why trust cannot be simply reduced to and replaced by a probability or risk measure.

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

¹¹ Moreover in fact with cognitive agents ($\text{Intend } y(\alpha, g) \implies (\text{Bel } y(\text{CanDo } y(\alpha)))$). If y believes that he cannot do α (or does not believe that he can do), he will not intend to do α .

¹² Of course, this deserves more elaboration and a specific work to introduce temporal specifications in this logics. This is out of the aims of this paper.

attribution) and the probability of having the appropriate conditions (opportunities and resources *external attribution*) for the performance and for its success, and of not having interferences and adversities (*external attribution*). Why this decomposition is important? Not only for cognitively grounding such a probability (which after all is 'subjective' i.e. mentally elaborated) - and this cognitive embedding is fundamental for relying, influencing, persuading, etc.-, but because:

- a) the agent trusting/delegating decision might be different with the same global probability or risk, depending on its composition;
- b) trust composition (internal Vs external) produces completely different intervention strategies: to manipulate the external variables (circumstances, infrastructures) is completely different than manipulating internal parameters.

Let's consider the first point. There might be different heuristics or different personalities with different propensity to delegate or not in case of a weak internal trust (subjective *trustworthiness*) even with the same global risk. For example, "I completely trust him but he cannot succeed, it is an impossible task!", or "The mission/task is not difficult, but I do not have enough trust in him". The problem is that - given the same global expectation - one agent might decide to trust/rely in one case but not in the other, or viceversa!

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

To be true, we should also consider the reciprocal influence between external and internal factors. When *x* trusts the internal powers of *y*, it also trusts his abilities to create positive opportunities for success, to perceive and react to the external problems. Viceversa, when *x* trusts the environment opportunities, this valuation could change the trust about *y* (*x* could think that *y* is not able to react to specific external problems).

Environmental and situational trust (which are claimed to be so crucial in electronic commerce and computer mediated interaction; see for ex. Castelfranchi e Tan, introduction in this volume; Rea, this volume) are aspects of the external trust. Is it important to stress (also Rea, this volume) that:

- *when the environment and the specific circumstances are safe and reliable, less trust in y (the contractor) is necessary for delegation (for ex. for transactions).*

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

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

3.3 Trust in strong delegation: adoption-based trust

Let's eventually arrive to social trust in strong delegation, which is its typical and strict sense in the social sciences. The mental attitude is the same (that is why is important to relate trust to *any* level of delegation), i.e. all previous beliefs hold, but there are some specific additional features. Strong Delegation is in fact based on *y*'s awareness and implicit or explicit agreement (compliance); it presupposes *goal-adoption* by *y*. Thus to trust *y* in this case means to trust his agreement and willingness to help/adopt (social commitment).¹³

Trusting motivation and morality of y

First of all, it is very important to analyze the beliefs about the motives of *y*. In particular, it is crucial to specify the beliefs about the adoptive (helping) attitude of *y* and its motives and persistence.

¹³ There are other cases of Adoption-based trust, i.e. of help, outside strong delegation and agreement. It is the case of weak delegation where *x* believe that *y* intends spontaneously help her. In all forms of adoption-based trust, beliefs about adoptivity and motives for adoption are particularly crucial.

8. **Motivation Belief.** x believes that y has some motives to help her (to adopt her goal), and that these motives will probably prevail -in case of conflict- on other motives, negative for her.

Notice that motives inducing to adoption are of several different kinds: from friendship to altruism, from morality to fear of sanctions, from exchange to common goal (cooperation).

This is why for example is important to have common culture, shared values, the same acknowledged authorities between trustor and trustee. The belief in shared values or in acknowledged authority [Con2] is an evidence and a basis for believing that y is sensible to certain motives and that they are important and prevailing.

In particular beliefs about y 's morality are relevant for trusting him ¹⁴. When there is a promise, not only y has an intention to do α , but he has such an intention (also) because he is "committed to" x to do α ; there is an (explicit or implicit) promise to do so which implies an interpersonal duty (x has some *rights* on y : to pretend, to complain, etc. -[Cas5]) and -in organisations, institution, societies - an obligation (derived from social norms) to do α (since he promised so to x).¹⁵ An additional trust is needed: the belief that y has been *sincere* (if he said that he intend to do it he really intends to do it) and that he is *honest/truthful* (if he has a commitment he will keep his promise; he will do what he ought to do).

On such a basis -of his adoptive disposition- x supports her beliefs that " y intends to do" and that "he will persist", and then the belief that he "will do".

Only this kind/level of social trust can be really "betrayed": if y is not aware of or didn't (at least implicitly) agree about x 's reliance and trust, he is not really "betraying" x .

Claims about the mind of the other: is trust a belief in the other's irrationality?

Of course, this kind of trust (the trust needed in promises, in contracts, in business, in organisations and collaboration) has been object of study in the social sciences. They correctly stress the relationship between sincerity, honesty (reputation), friendliness and trust.

However, sometimes this has not been formulated in a very linear way; especially under the perspective of game theory and within the framework of the Prisoner Dilemma that strongly influenced all the problem of defection, cheating, and social dilemma.

Consider for example the definition of trust proposed by [Gam] in his interdisciplinary discussion on trust.

"When I say that I trust y , I mean that *I believe that, put on test, y would act in a way favourable to me, even though this choice would not be the most convenient for him at that moment*".¹⁶

So formulated, (considering subjective rationality) *trust is the belief that y will choose and will behave in a non-rational way!* How might he otherwise choose what is perceived as less convenient? This is the usual dilemma in the PD game: the only rational move is to defect.

Since trust is one of the pillars of society (no social exchange, alliance, cooperation, institution, group, is possible without trust), should we conclude that the entire society is grounded on the irrationality of the agents: either the irrationality of y , or the irrationality of x in believing that y will act irrationally, against his better interest!

As usual in arguments and models inspired by rational decision theory or game theory, with rationality also "selfishness" and "economic motives" (utility, profit) are smuggled [Cas4].

When x trusts y in strong delegation (goal-adoption and social commitment by y) x is not assuming that he -by not defeating her- acts irrationally, i.e. against his interests. Perhaps he acts "economically irrationally" (i.e. sacrificing his economic goals); perhaps he acts in an unselfish way, preferring to his selfish goals some altruistic or pro-social or normative motive; but he is not irrational because he is just following his subjective preferences and motives, including friendship, or love, or norms, or honesty, etc.

¹⁴ 'Morality' and 'shared values' must be considered in a relative, cultural way. Also a thief or a 'mafioso' trusts his accomplices as for their 'morality' (conspiracy of silence, fidelity, honour, etc.).

¹⁵ Quite similar is the case in which I trust y because what I expect from him is obligatory for him: there is a norm, a law prescribing that behaviour. In this case I bet on y 's normative motives, even if there might not be any agreement or compliance (weak delegation). For example I trust/believe that the other drivers will stop at the red light, and I rely on this.

¹⁶ The same holds in the definition of Lieberman [Lie]: trust is "a belief or expectation that the parties involved in the agreement will actually do what they have agreed to do; they will fulfill their commitments not only when it is obviously advantageous to do so, but *even when it may be disadvantageous*, when they must sacrifice some immediate gain".

Thus when x trusts y , x is just assuming that other motivations will prevail over his economic interests or other selfish goals.

Assertion 10: *Trust is a theory and an expectation about the kind of motivations the agent is endowed with, and about which will be the prevailing motivations in case of conflict.*

x not only believes that y will intend and persist (and then he will do) but x believes that y will persist *because of certain motives* of his that are more important than other motives inducing him to defection and betrayal. And these motives are already there -in y 's mind and in their agreement- x has not to find new incentives, to think of additional prizes or of possible punishments. If x is doing so (for ex. by promising or threatening) x does not really trust y (yet).

3.4 Internal trust, unharfulness and adoptive attitude (goodwill)

The theory of the internally attributed trust is important also for clarifying and modelling that additional nucleus of trust that we named "unharfulness" or "adoptivity" and rapidly putted aside. In fact we precisely refer to a sense of safety, an feeling that "there is nothing to worry about as from y ", no danger, no suspect, no hostility, and more than this the idea that the other will help, is well disposed, will care of our interest, will adopt them. The belief that " y would act in a way favourable to me" (Gambetta, cit.).

Now this nucleus is part of the *internal* trust, and precisely refers to the social aspect of it. We might say that the 'trust in' y distinguishes between y 's general mental attitudes relevant for a reliable action, and his 'social' and more precisely 'adoptive' or pro-social aspect, i.e. y 's disposition towards me (or people in general).

The first part of internal trust is important in any kind of delegation but especially in weak-delegation where y might ignore that we are exploiting his action, and does not usually care of our interests or goals. In this situation what is important is y 's personal commitment, his will, his persistence, his ability, etc. On the contrary if our reliance/delegation is based on the fact that y 's takes into account our goals and possibly favour them (goal-adoption, help - [Cas3, Mic] or at least avoid to damage them (*collaborative coordination* [Cas7], *passive goal-adoption* - [Cas8], or even strongly it is based on y 's social-commitment (promises, role, etc.) towards us, in this case what we believe about his disposition towards us and our interest is very crucial and is an relevant basis of our trust. In fact Strong Delegation presupposes y 's goal-adoption, her/his acceptance of my reliance, her/his adhesion to my (implicit or explicit) request or her/his spontaneous help. And x believes that his bet will be successful thanks to y 's acceptance and adoption and willingness to be helpful. Moreover, x trust on the persistence of the collaborative attitude or pro-social and on its prevalence on interfering motives. This is what [Bon] interestingly proposes to identify as y 's 'goodwill' (although she aims to put in it all the different aspects of the internal trust in cognitive agent, that we prefer to distinguish).

In sum, if y 's goal-adoption is the basis of x 's delegation¹⁷, then x counts on y 's adoption of her/his goal, not simply on y 's action or intention. x believes/trusts that y will do a for x ! More precisely, trust that y will do is implied and supported by the trust that y will do *for x* .

This is why *the beliefs about the social disposition of y are crucial* (although we know that goal adoption can be for several kinds of motives, from selfish to altruistic, cit).

So, we claim that internal attribution distinguish among three areas:

- the capacity of y 's (skilful, know how, careful and accurate, self-confidence, ..);
- the non-social generic motivational aspects (intends, persists, is seriously engaged, effort, ..);
- the social attitudes, this basically consist in the belief/feeling that there is a pro-attitude [Tuo], a 'goodwill' towards (also) us and that there are no conflictual (anti-social) motives or, in any case, the adoptive attitude (for whatever motivation, [Con1]) will prevail. The weaker form of this component of trust is the belief of 'unharfulness': there are no danger, no hostility, no defeating attitudes, no antipathy, etc.

Doubts, suspects, can separately affect these three facets of our internal trust 'in' y .

He is not so expert, or so skilled, or so careful, ...or he is not enough smart or reactive to recognise and exploit opportunities or to cope with interferences,; he is quite voluble, or not enough engaged and putting effort, or he has conflict of preferences: he will and will not at the same time,...

¹⁷ This is the case both in weak-delegation based on spontaneous help without agreement, and -obviously- in strong delegation.

there is some unconscious hostility, some antipathy, he does not care so much of me, he is quite selfish and egocentric,

In this framework it is quite clear why we trust friends. First we believe that as friends they want our good, they want to help us; thus they both will adopt our request and will keep their promise. Moreover, they do not have reasons for damaging us or for hiddenly aggress against us. Even if there is some conflict, some selfish interest against us, friendship will be more important for them. We rely on the *motivational strength* of friendship.

3.5 The greatest the autonomy the deepest the trust. Open delegation and delegation of control

With cognitive, autonomous agents it is possible to have "open delegation". In Open delegation [Cas2] *x* delegates to *y* a goal to be achieved rather than a specific performance. *y* has to "bring it about that *g*"; he should find a correct plan, choose, plan, adapt, and so on. *x* either ignores or does not specify the necessary action or plan. *y* is more autonomous, and *x* must trust also *y*'s cognitive ability in choosing and planning; *x* is in fact depending not only on *y*'s resources and practical abilities, but also on *y*'s problem-solving capacity and knowledge: he must be *competent* on the delegated problem. In social trust we are really betting on *y*'s mind.

The deepest level of trust with a fully autonomous agent is the delegation of or *the renunciation to the monitoring and control*. *x* is so sure that *y* will do what *x* expects (for ex. what he promised) that *x* does not check up or inspect. In fact when we monitor or inspect somebody who is doing something we need, he can complain with us and say: "*this means that you don't trust me!!*". Of course, renouncing the control increases the risk, since it increases the possibility that *x* is deceived and delays possible repairs or protections.

4. Trust as a three party relationship

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

In our view this is correct only if one adopts a quite limited view of trust in terms of beliefs relative to the character or friendliness, etc. of the trustee (delegated agent). In fact in these cases (contracts, organisations) we just deal with *a more complex and specific kind of trust*. But trust is always crucial.

We put a contract in place only because we believe that the agent will not violate the contract, and this is precisely "trust". We base this trust in the contractor (the belief that s/he will do what promised) either on the belief that s/he is a moral person and keeps her/his promises, or on the belief that s/he worries about law and punishment.

To be more clear, this level of trust is a three party relationship: it is a relation between a client *x*, a contractor *y* and the authority *A*. And there are three trust sub-relations in it:

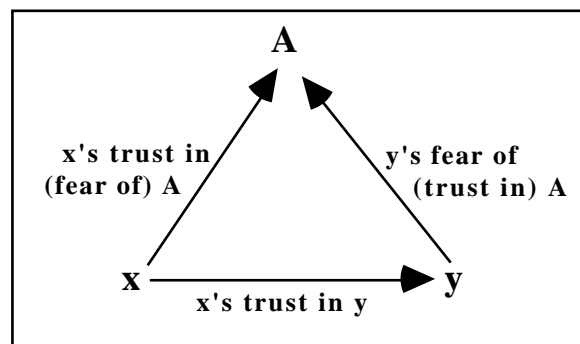


Figure 1

$\text{Trust}(X, Y, \tau) = B_X \text{Ability}_Y(\tau) \wedge B_X \text{WillDo}_Y(\tau, g)$
 $\text{Trust}(X, A, \tau) = B_X \text{Ability}_A(\tau) \wedge B_X \text{WillDo}_A(\tau, g')$
 $\text{Trust}(Y, A, \tau) = B_Y \text{Ability}_A(\tau) \wedge B_Y \text{WillDo}_A(\tau, g')$

where τ is the task that Y must perform for X; τ' the task that A must perform for X towards Y, i.e. check, supervision, guarantee, punishment, etc. (Of course, Y's trust about τ' is strange, since τ' is a danger for him - see later). More precisely and importantly in X's mind there is a belief and a goal (thus an expectation) about this trust of Y in A:

$B_X(\text{Trust}(Y, A, \tau)) \wedge \text{Goal}_X(\text{Trust}(Y, A, \tau))$.

And this expectation is the new level of X's trust in the contractor.

X trusts Y by believing that Y will do what promised because of her/his honesty or because of her/his respect/fear toward A. In other words, X relies on a form of paradoxical trust of Y in A: X believes that Y believes that A is able to control, to punish, etc.

Of course, normally a contract is bilateral and symmetric, thus the point of view of Y's should be added, and his trust in X and in A as for monitoring X.

Notice that Y's beliefs about A are precisely Y's trust in the authority when s/he is the client, while, when Y is the contractor, the same beliefs are the bases of her/his respect/fear toward A.

(Positive) Trust Vs Aversive Trust: The common core of trust and fear

There is a paradoxical but effective form of trust, which is trust in threats and threatening agents: fear as a form of trust. The basic core of the two attitudes is in fact the same. This is important since what in a given circumstance appears as worry and fear produces in another circumstance a normal trust. As we just said, in a contract X trusts Y because X trusts the authority's capacity of punishing Y in case of violation, and because X trusts Y's fear of authority and punishments. But *Y's fear of authority is precisely his trust in authority* as able to check and punish X (and viceversa): the same evaluations of the authority are at the same time trust in it and wonderment / awe/ fear depending on the point of view.

The core components of these two mental attitudes of positive and negative respect are the same, and we already identified them:

$B_X \text{Ability}_Y(\tau) \wedge B_X \text{WillDo}_Y(\tau, g)$

We presented them as the two basic beliefs of (positive) trust (Cfr. fig 1), but this in fact only depends on X's goal: if X has the goal g

($\text{Goal}_X g$)

we have a positive expectation [Cas9] and then (positive) Trust. On the contrary if X does not want what she expects

($\text{Goal}_X \neg g$)

we have negative expectation and aversive trust.

In this case, for X Y's power of doing τ is a threat, is a negative power, and Y's propensity (ex. willingness) to do τ is a bad-will.

The presence of the positive goal makes the two beliefs the core of 'reliability' and Trust; the presence of the negative goal makes them the core of Y's frightens.

Reliability and trust is what makes effective promises (in fact, what is promised is a goal of X). Frightens is what makes threats effective (since what is threatened is the opposite of a goal of X).¹⁸

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

As we said these more complex kinds of trust are just more rich *specifications of the reasons for Y's doing what we expect*: reasons for Y's predictability (WillDo) which is based on her/his willingness (IntedToDo); and reasons for her/his willingness (s/he will do τ , either because of

¹⁸ When we say "it promises to rain" or "it threatens to rain" our beliefs are the same, only our goal (and then our attitude) changes: in the first case we want rain, in the second we dislike it.

her/his selfish interest, or because of her/his friendliness, or because of her/his honesty, or because of her/his fear of punishment: several different bases of trust).

4.1 Increasing trust: From Intentions to Contracts

What we have just described are not only different kinds and different bases of trust. They can be conceived also as different levels/degrees of social trust and *additional* supports for trust. We mean that one basis does not necessary eliminate the other but can supplement it or replace it when is not sufficient. If I do not trust enough in your personal persistence I can trust in your keeping your promises, and if this is not enough (or is not there) I can trust in your respecting the laws or in your worrying about punishments.

We consider these 'motivations' and these 'commitments' not all equivalent: some are stronger or more cogent than others. As we claimed in [Cas10]

This more cogent and normative nature of S-Commitment explains why abandoning a Joint Intention or plan, a coalition or a team is not so simple as dropping a private Intention. This is not because the dropping agent must inform her partners -behaviour that sometimes is even irrational-, but precisely because Joint Intentions, team work, coalitions (and what we will call Collective-Commitments) imply S-Commitments among the members and between the member and her group. In fact, one cannot exit a S-Commitment in the same way one can exit an I-Commitment. Consequences (and thus utilities taken into account in the decision) are quite different because in exiting S-Commitments one violates obligations, frustrate expectations and rights she created. We could not trust in teams and coalitions and cooperate with each others if the stability of reciprocal and collective Commitments was just like the stability of I-Commitments (Intentions).

Let us analyse more carefully this point, by comparing 5 scenarios of delegation:

a) Intention ascription

X is weakly delegating Y a task (let say to raise his arm and stop the bus) on the basis just on the hypothetical ascription to Y of an intention (he intends to stop the bus in order to take the bus).

There are two problem in this kind of situation:

- the *ascription of the intention* is just based on abduction and defeasable inferences, and to rely on this is quite risky (we can do this when the situation is very clear and very constrained by a script, like at the bus stop);
- this is just *a private intention and a personal commitment* to a given action; Y can change his private mind as he likes; he has no social binds in doing this.

b) Intention declaration

X is weakly delegating Y a task (to raise his arm and stop the bus) on the basis not only of Y's situation and behaviour (the current script) but also or just on the basis of a declaration of intention by Y. In this case both the previous problems are a bit better:

- the *ascription of the intention* is more safe and reliable (excluding deception that on the other side would introduce normative aspects that we deserve for more advanced scenarios);
- now *Y knows that X knows about his intention* and about his declaring his intention; there is no promise and no social commitment to X, but at least changing his mind Y should care of X evaluation about his coherence or sincerity or fickleness; thus he will a bit more bound to his declared intention, and X can rely a bit more safely on it.

In other terms X's degree of trust can increase because of:

- either a larger number of evidences;
- or a larger number of motives and reasons for Y doing ;
- or the stronger value of the involved goals/motives of Y.

There is an implicit law here:

the greater the number of independent motives for doing τ , and the greater their importance or values the greater the probability of doing τ .

c) Promises

Promises are stronger than simple declaration or knowledge of the intention of another agent. Promises create what we called a Social-Commitment, which is a right producing act, and determine rights for X and duties/obligations for Y. We claim that this is independent on laws, authority,

punishment. It is just at the micro level, as inter-personal, direct relation (not mediated by a third party, be it a group an authority, etc.).

The very act of committing oneself to someone else is a "rights-producing" act: before the S-Commitment, before the "promise", y has no rights over x , y is not entitled (by x) to exact this action. After the S-Commitment it exists such a new and crucial social relation: y has some rights on x , she is entitled by the very act of Commitment on x 's part. So, the notion of S-Commitment is well defined only if it implies these other relations:

- y is entitled (to control, to exact/require, to complain/protest);
- x is in debt to y ;
- x acknowledges to be in debt to y and y 's rights.

In other terms, x cannot protest (or better *he is committed to not protesting*) if y protests (exacts, etc.).

One should introduce a relation of "entitlement" between x and y meaning that y has the rights of controlling a , of exacting a , of protesting (and punishing), in other words, x is S-Committed to y to not oppose to these rights of y (in such a way, x "acknowledges" these rights of y). [Cas10]

If Y changes his mind he is disappointing X 's entitled expectations and frustrating X 's rights. He must expect and undergo X 's disappointment, hostility and protests. He is probably violating shared values (since he agreed about X 's expectations and rights) and then is exposed to internal bad feelings like shame and guilt. Probably he does not like all this. This means that *there are additional goals/motives that create incentives for persisting in the intention*. X can reasonably have more trust.

Notice that also the declaration is more constraining in promises: to lie is more heavy.

d) Promises with witness and oaths

Even stronger is a promise in front of a witness, or a oaths (which is in front of god). In fact, there are additional bad consequences in case of violation. Y would jeopardise his reputation (with very bad potential consequences; see [Cas11]) receiving a bad evaluation also from the witness; or he bad in front of god eliciting his punishment.

Thus if I do not trust what you say I will ask you to promise this; and if I do not trust your promise I ask you to promise in front of other people or to oath about this. If I worry that you might negate that you promise I will ask for writing and sign something. And so on.

e) Contracts

Even public promises might be not enough and we may proceed in adding binds to binds in order to make Y more predictable and more reliable. In particular we might exploit more the third party. We can have a group [Sin], an authority able to issue norms (defending rights and creating obligations), to control violation, to punish violators. Of course this authority or reference group must be shared and acknowledge, and as we said trusted by both X and Y . Thus we got an additional problem of trust. However, now Y has additional reasons for keeping its commitment, and X 's degree of trust is higher.

Noticed that all these additional beliefs about Y are specific kinds or facets of trust in Y : X trusts that Y is respectful of norms, or that Y fears punishments; X trust in Y 's honesty or shame or ambition of good reputation, etc.

A bit more precisely:

- the stronger the Y 's motive for doing and then the stronger his commitment to ;
- the larger the number of those motives;
- and the stronger X 's beliefs about this;

the stronger it will be X 's trust in Y as for doing

5. Degrees of Trust: a principled quantification of Trust

The idea that trust is scalable is usual (in common sense, in social sciences, in AI [Sni, Mar]). However, since no real definition and cognitive characterisation of trust is given the quantification of trust is quite *ad hoc* and arbitrary, and the introduction of this notion or predicate is semantically

empty.¹⁹ On the contrary we claim that there is a strong coherence between the cognitive definition of trust, its mental ingredients, and, on the one side, its value, on the other side, its social functions and its affective aspects (we will not examine here). More precisely the latter are based on the former.

Here we will ground the degree of trust of x in y , on the cognitive components of x 's mental state of trust. More precisely we claim that *the degree of trust is a function of the subjective certainty of the pertinent beliefs*. In section 6. we will use the degree of trust to formalise a rational basis for the decision of relying and betting on y . Also in this case we will claim that the "quantitative" aspect of another basic ingredient is relevant: *the value or importance or utility of the goal g* , will obviously enter the evaluation of the risk, and will also modify the required threshold for trusting. In sum,

- *the quantitative dimensions of trust are based on the quantitative dimensions of its cognitive constituents.*

For us trust is not an arbitrary index just with an operational importance, without a real content, but it is based on the subjective certainty of the pertinent beliefs.

5.1 Trust in beliefs and trust in action and delegation

The solution we propose is not an *ad hoc* solution, just to ground some degree of trust. It instantiates a general claim. [Pea] points out the relation between the level of confidence in a belief and the likelihood of a person taking action based on the belief: "Think of the person who makes a true statement based on adequate reasons, but does not feel confident that it is true. Obviously, *he is much less likely to act on it, and, in the extreme case of lack of confidence, would not act on it*" (p. 15) (We stressed the terms clearly related to theory of trust).

"It is commonly accepted that people behave in accordance with their knowledge" (Notice that this is precisely our definition of a 'cognitive agent'! but it would be better to use 'beliefs').

"*The more certain the knowledge then the more likely, more rapid and more reliable is the response.* If a person strongly believes something to be correct which is, in fact, incorrect, then the performance of the tasks which rely on this erroneous belief or misinformation will likewise be in error - even though the response may be executed rapidly and with confidence." [Hun, p.8]. Thus under our foundation of the degree of trust there is a general principle:

- Agents act depending on what they believe, i.e. *relying on* their beliefs. And they act on the basis of the degree of reliability and certainty they attribute to their beliefs. In other words, trust/confidence in an action or plan (reasons to choose it and expectations of success) is grounded on and derives from trust-confidence in the related beliefs!!!

The case of trust in delegated tools or agents is just a consequence of this general principle of action in cognitive agents. Also beliefs are something one bets and risks on, when he decides of basing his action on them. And chosen actions too are something one bets, relies, count on and depend upon. We trust our beliefs, we trust our actions, we trust delegated tools and agents. In an uncertain world any single action would be impossible without some form of trust [Luh].

5.2 A belief-based Degree of trust

Let's call the degree of trust of X in Y about τ : $\text{DoT}_{XY\tau}$ ($0 \leq \text{DoT}_{XY\tau} \leq 1$). Given that we postulate that the degree of trust is a function of the "strength" of the trusting beliefs, i.e. of their *credibility* (expressing both the subjective probability of the fact and trust in the belief): the greater X 's belief in Y 's competence and performance, the greater X 's trust in Y .

¹⁹ Williamson [Wil] for example claims that 'trust' is an empty and superfluous notion -used by sociologists just for rethorics- since it is simply reducible to subjective probability/risk.

$$DoT_{XY\tau} = DoC_X[Opp_Y(\cdot, g)] * DoC_X[Ability_Y(\cdot)] * DoC_X[WillDo_Y(\cdot, g)]$$

where:

- $DoC_X[Opp_Y(\cdot, g)]$, is the degree of credibility of X's beliefs about the Y's opportunity of performing to realize g;
- $DoC_X[Ability_Y(\cdot)]$, the degree of credibility of X's beliefs about the Y's ability/competence to perform ;
- $DoC_X[WillDo_Y(\cdot, g)]$, the degree of credibility of X's beliefs about the Y's actual performance;

$$DoC_X[WillDo_Y(\cdot, g)] = DoC_X[Intend_Y(\cdot, g)] * DoC_X[Persist_Y(\cdot, g)]$$

(given that Y is a *cognitive agent*)

We assume that the various credibility degrees are independent from each other.

6. To trust or not to trust: degrees of trust and decision to trust

In any circumstance, an agent X endowed with a given goal, has three main choices:

- i) to try to achieve the goal by itself;
- ii) to delegate the achievement of that goal to another agent Y;
- iii) to do nothing (relatively to this goal), renouncing.

So we should consider the following abstract scenario (Fig. 2) where we call:

- $U(X)$, the agent X's utility function, and specifically:
- $U(X)_{p+}$, the utility of the X's success performance;
- $U(X)_{p-}$, the utility of the X's failure performance;
- $U(X)_{d+}$ the utility of a successful delegation (the utility due to the success of the delegated action);
- $U(X)_{d-}$ the utility of a failure delegation (the damage due to the failure of the delegated action);
- $U(X)_0$ the utility of to do nothing.

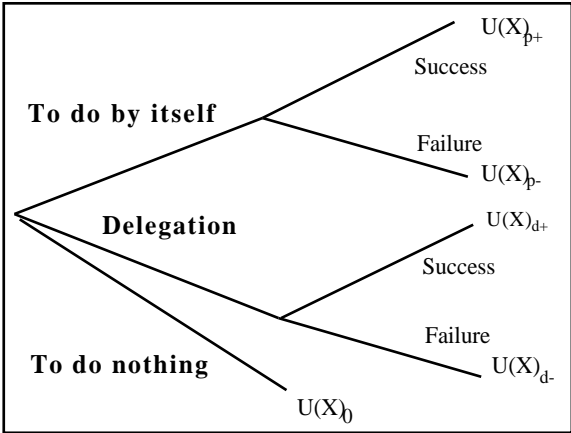


Figure 2

However, for sake of brevity, we will consider a simplified scenario:

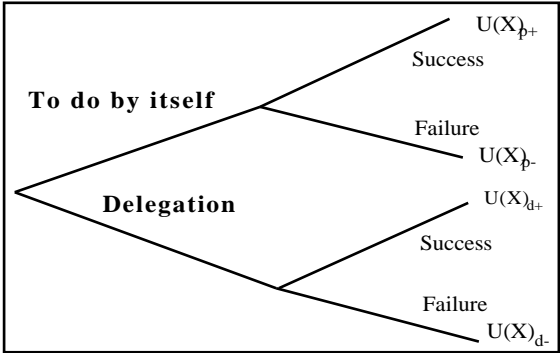


Figure 3

In the above scenario (Figure 3), in order to delegate we must have:

$$\text{DoT}_{XY} * U(X)_{d^+} + (1 - \text{DoT}_{XY}) U(X)_{d^-} > \text{DoT}_{XX} * U(X)_{p^+} + (1 - \text{DoT}_{XX}) U(X)_{p^-}$$

where DoT_{XY} is the *selftrust* of X about .²⁰

Then, we obtain

$$\text{DoT}_{XY} > \text{DoT}_{XX} * A + B \quad (1)$$

where:

$$A = (U(X)_{p^+} - U(X)_{p^-}) / (U(X)_{d^+} - U(X)_{d^-})$$

$$B = (U(X)_{p^-} - U(X)_{d^-}) / (U(X)_{d^+} - U(X)_{d^-})$$

Let us consider now, the two terms A and B separately.

As for the term A, if $U(X)_{p^+} - U(X)_{p^-} > U(X)_{d^+} - U(X)_{d^-}$ then $A * \text{DoT}_{XX} > \text{DoT}_{XX}$

i.e. if the difference between the utility of the success and the utility of the failure in delegation is smaller than the difference between the utility of the success and the utility of the failure in non delegation, then (for the term A) in order to delegate

• *the trust of X in Y must be bigger than the selftrust of X (about)*.

Vice versa, if $U(X)_{p^+} - U(X)_{p^-} < U(X)_{d^+} - U(X)_{d^-}$ then $A * \text{DoT}_{XX} < \text{DoT}_{XX}$

i.e., if the difference between the utility of the success and the utility of the failure in delegation is bigger than the difference between the utility of the success and the utility of the failure in non delegation, then (for the term A) in order to delegate

• *the trust of X in Y could be smaller than the selftrust of X (about)*.

So, it is possible also to delegate to people which I trust less than myself.

Considering now the term B,

if $U(X)_{p^-} - U(X)_{d^-} > 0$, then a positive term is added to the A: $A + B > A$,

i.e., if the utility of the failure in case of non-delegating is bigger than the utility of the failure in case of delegation, then - in order to delegate - the trust of X in Y about must be greater than in the case in which the right part of (1) is constituted by A alone.

Viceversa, if $U(X)_{p^-} - U(X)_{d^-} < 0$, then $A + B < A$,

i.e., if the utility of the failure in case of non-delegating is smaller than the utility of the failure in case of delegation, then - in order to delegate - the trust of X in Y about must be smaller than in the case in which the right part of (1) is constituted by just A alone.²¹

Since $\text{DoT}_{XY} \leq 1$, from the (1) we can obtain:

$$\text{DoT}_{XX} < (U(X)_{d^+} - U(X)_{p^-}) / (U(X)_{p^+} - U(X)_{p^-}) \quad (2)$$

From (2) we can say that, to delegate X to Y the task , as the selftrust (DoT_{XX}) grows, the difference between the utility of the success in delegation and the utility of the failure in the non delegation should be reduced.

Moreover (to delegate), as the selftrust (DoT_{XX}) grows, it must reduce the difference between the utility of the success and of the failure in non delegation.

Because $\text{DoT}_{XX} \geq 0$, from (2) we obtain:

$$U(X)_{d^+} > U(X)_{p^-} \quad (3)$$

(consider that for definition we have $U(X)_{p^+} > U(X)_{p^-}$).

In practice, for delegating, a necessary (but not sufficient) condition is that the utility of the success in delegation is greater than the utility of the failure in the non delegation.

²⁰ More precisely, we have: $U(X)_{p^+} = \text{Value}(g) + \text{Cost} [\text{Performance}(X \ t)]$,

$U(X)_{p^-} = \text{Cost} [\text{Performance}(X \ t)] + \text{Additional Damage for failure}$

$U(X)_{d^+} = \text{Value}(g) + \text{Cost} [\text{Delegation}(X \ Y \ t)]$,

$U(X)_{d^-} = \text{Cost} [\text{Delegation}(X \ Y \ t)] + \text{Additional Damage for failure}$

where is supposed that it is possible to attribute a quantitative value (importance) to the goals and where the costs of the actions (delegation and performance) is supposed to be negative.

²¹ Both for A and B there is a normalization factor ($U(X)_{d^+} - U(X)_{d^-}$): the more its value increases, the more the importance of the terms is reduced.

6.1 Positive trust is not enough: A variable threshold for risk acceptance/avoidance

As we saw, *the decision to trust is based on some positive trust*, i.e. on some evaluation and expectation (beliefs) about the capability and willingness of the trustee and the probability of success.

First, those beliefs can be well justified, warranted and based on reasons. This represents the “rational” (reasons based) part of the trust in y . But they can also be not really warranted, not based on evidences, then quite irrational, faithful. We call this part of the trust in y : “faith”.²²

Notice that irrationality in trust decision can derive from these unjustified beliefs, i.e. on the ratio of mere faith (see section 7.2).

Second, *positive trust is not enough* for accounting for the decision to trust/delegate. We do not distinguish in this paper the different role or impact of the rational and irrational part of our trust or positive expectations about y action: the entire positive trust (reason-based + faithful) is necessary and contributes to the DoT: its sum should be greater than discouraging factors. We do not go deeply in this distinction (a part from the problem of rational Vs irrational trust) also because we are interested here in the additional fact that these (grounded or ungrounded) positive expectation is not enough for explaining the *decision/act* of trusting. In fact, another aspect is necessarily involved in this decision, as we said discussing uncertainty, risk, and bet. The decision to trust/delegate necessarily implies *the acceptance of some perceived risk*. A trusting agent is a risk-acceptant agent. Trust is never certainty: always it remains some uncertainty (ignorance) and some probability of failure, and the agent must accept this and to run such a risk.

Thus a fundamental component of our decision to trust y , is our acceptance and felt exposition to a risk. Risk is represented in previous quantification of DoT and in criteria for decision. However, we believe that this is not enough. A specific risk policy seems necessary to trust and bet, and we should explicitly capture this aspect.

The equation (1) - that basically follows classical decision theory - introduces the degree of trust instead of simple probability factor. In this way, it permits to evaluate when to delegate rather than to do by itself in a rigid rational way. The importance of this equation is to establish what decision branch is the best on the basis of both the relative (success and failure) utilities for each branch and the probability (trust based) of them. In this equation no factor can play a role independently from the others. Unfortunately in several situations and contexts, not just for the human decisors but -we think- also for good artificial decisors, it should be important to consider the absolute values of some parameter independently from the values of the others. This fact suggests to introduce some saturation-based mechanism to influence the decision, some threshold.

For example it is possible that the value of the damage *per se* (in case of failure) is too high to choose a given decision branch, and this independently either from the probability of the failure (even if it is very low) or from the possible payoff (even if it is very high). In other words, that danger might seem to the agent an intolerable risk.

In this paragraph we analyze (just in a qualitative way) different possible threshold factors that must play an additional role to choose among alternatives like in figure 3.

First, let us assume that each choice implies a given failure probability as perceived by X (and let's call this 'hazard' or 'danger'), and a given 'threat' or 'damage': i.e. a negative utility due to both the failure (the cost of a wasted activity and a missed reward) plus possible additional damages.²³

Second, we assume that X is disposed to accept a maximum hazard (H_{max}) in its choices, in a given domain and situation. In other words, *there is a 'hazard' threshold over which X is not disposed to pursue that choice*.

²² To be more precise, non-rational blind trust is close to faith. Faith is more than trust without evidences, it is trust without the need for and the search for evidences.

²³ Thus we will use here the term 'risk' as the result of the entity of losses (damage or threat) and of its probability (hazard or danger). Risk theory [Kap] calculates the risk as the product of uncertainty (subjective probability) and damage; other authors propose -for the objective risk- the product of frequency and magnitude of the danger. We are interested in the subjective dimension, so risk should be $U \cdot pb$; in our terminology hazard * damage. (Common sense would prefer to call 'risk' the probability, and 'danger' the global result of probability and damage).

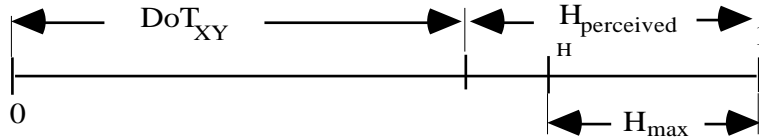


Figure 4

We are considering the case of delegation branch (Dot_{XY} , $U(X)_{d^-}$, $U(X)_{d^+}$), but the same concepts are valid in the case of X's performance (substituting Dot_{XX} , $U(X)_{p^-}$, $U(X)_{p^+}$). In figure 4 we have:

$H_{\text{perceived}}$ is the failure hazard perceived by X;

$H_{\text{max}} = (1 - R)$ is the maximum failure hazard acceptable by X;

R is the hazard threshold.

To choose a given path it is necessary that:

$$\text{Dot}_{XY} \geq H = (1 - H_{\text{max}})$$

We claim that such a threshold can vary, not only from one agent to another (personality) but also depending on several factors in the same agent. In particular we claim that the acceptable hazard varies with the importance of the *threat-damage* and with the expected reward. In other words,

$$H \text{ (where } 0 \leq H \leq 1) \text{ is a function of both } (U(X)_{d^-}) \text{ and } (U(X)_{d^+}): H = f(U(X)_{d^-}, U(X)_{d^+})$$

More precisely: the greater the damage ($U(X)_{d^-}$) more it grows H ; while the greater the utility ($U(X)_{d^+}$) the more H is reduced.

Moreover, we may introduce also an 'acceptable damage' threshold d ²⁴. The function H is such that when $U(X)_{d^-}$ is equal (or lesser) than d then H is equal to 1 (in practice, that choice is impossible). For each agent both d and H can assume different values.

One might also have one single dimension and threshold for *risk* (by using the formula 'damage * hazard'). However, we claim that there could be different heuristics for coping with risk (for sure this is true for human agents). For us, a great damage with a small probability and a small damage with a high probability, do not represent two equivalent risks. They can lead to different decisions, they can differently pass or not the threshold.

To resume in the case of delegation branch (it is sufficient to substitute $U(X)_{d^-}$ with $U(X)_{p^-}$, $U(X)_{d^+}$ with $U(X)_{p^+}$, to obtain the case of X's performance branch) we have:

$$H = f(U(X)_{d^-}, U(X)_{d^+}) \text{ and in particular } H = 1 \text{ when } U(X)_{d^-} \leq d.$$

In other words, we assume that *there is a risk threshold* -more precisely a *hazard* threshold depending also on a *damage* threshold- under which the agent refuses a given choice even if the equation (1) suggests that choice as the best. It might be that a choice is convenient (and the best) as for the ratio between possible payoff, costs and risk, but that the risk *per se* is too high for that agent in that situation. Let us consider an example (Figure 5):

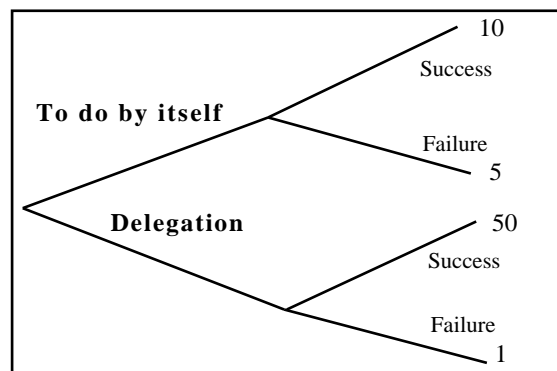


Figure 5

²⁴ We might introduce a minimal acceptable value for $U(X)_{d^+}$ (p , *payoff threshold*) under which a choice would be regarded as an unacceptable.

given $U(X)_p^+ = 10$, $U(X)_p^- = 5$, $U(X)_d^+ = 50$, $U(X)_d^- = 1$, and $\text{Dot}_{XY} = \text{Dot}_{XX} = 0.7$, the equation (1) is satisfied: $0.70 > (0.70 * 5/49) + (4/49) = 0.15$. So on the basis of this equation the agent X should delegate the task to the agent Y.

However, suppose that the maximum acceptable damage for X is $d = 4$ (the damage grows as the $U(X)_d^-$ is reduced) then the choice to delegate is stopped from the saturation effect.²⁵

7. When trust is too few or too much

Trust is not always rational or always adaptive and profitable. Let's see when it is rational or irrational, and when it is not useful although well grounded.

7.1 Rational trust

In our view trust can be rational and can support rational decisions²⁶. Trust as attitude (*core Trust*) is epistemically rational when is reason-based. When it is based on well motivated evidences and on good inferences, when its constitutive beliefs are well grounded (their credibility is correctly based on external and internal credible sources); when the evaluation is realistic and the esteem is justified, not mere faith.

The decision/action of trusting is rational when is based on a epistemically rational attitude and on a sufficient degree relative to the perceived risk. If my expectation is well grounded and the degree of trust exceeds the perceived risk, my decision to trust is subjectively rational.

To trust is indeed irrational either when the accepted risk is too high (relative to the degree of trust), or when trust is not based on good evidences, is not well supported. Either the faith component (unwarranted expectations) or the risk acceptance (blind trust) are too high.²⁷

Over-confidence and over-diffidence

Trust is not always good -also in cooperation and organisation. It can be dangerous both for the individual and for the organisation. In fact the consequences of over-confidence (the excess of trust) at the individual level are: reduced control actions; additional risks; non careful and non accurate action; distraction; delay in repair; possible partial or total failure, or additional cost for recovering. The same is true in collective activity. But, what does it mean 'over-confidence' i.e. excess of trust? In our model it means that x accepts too much risk or too much ignorance, or is not accurate in her evaluations. Noticed that there cannot be too much positive trust, esteem of y. It can

²⁵ It is possible that all the branches in the decision scenario are in a situation of saturation ($r = 1$). What choice the agent should decide? In these cases there could be several different possibilities. Let us consider the scenario in figure y. We could have at least four possibilities:

- saturation due to d for branch "to do by itself"; saturation due to a for branch "delegation";
- saturation due to a for branch "to do by itself"; saturation due to d for branch "delegation";
- saturation due to a for branch "to do by itself"; saturation due to a for branch "delegation";
- saturation due to d for branch "to do by itself"; saturation due to d for branch "delegation".

In the cases (a) and (b) the choice will always be the minimum damage. In the cases (c) if $(a - U(X)_d^+) > (a - U(X)_p^+)$ then the choice will be "to do by itself" and viceversa in the opposite case. In the case (d) if $(U(X)_d^+ - d) > (U(X)_p^+ - d)$ then the choice will be "to do by itself" and viceversa in the opposite case.

In the cases (c) and (d) if $(a - U(X)_d^+) = (a - U(X)_p^+)$ and $(U(X)_d^+ - d) = (U(X)_p^+ - d)$ then will be the equation (1) that will decide what is the right choice.

²⁶ We disagree with restrictive definitions of trust usual in Game Theory. Consider for example this definition:

"In general, we say that a person 'trusts someone to do X' if she acts on the expectation that he will do X when two conditions obtain: both know that if he fails to do X she would have done better to act otherwise, and her acting in the way she does gives him a selfish reason not to do X." (Bacharach and Gambetta in this book)

In this definition we recognise the 'Prisoner Dilemma syndrome' that gives an artificially limited and quite pessimistic view of social interaction. In fact, it is true that by trusting the other y makes herself 'vulnerable'; in other terms, she gives the other the opportunity to damage her. However, not necessarily she gives him a motive, a reason for damaging her.

²⁷ Rational trust can be based not only on reasons and reasoning, on explicit evaluations and beliefs, but also on simple learning and experience. For example the prediction of the event or result can not be based on some understanding of the process or some model of it, but just based on repeated experiences and associations.

be not well grounded and then bad placed: the actual risk is greater than the subjective one. Positive evaluation on y (trust in y) can be too much only in the sense that it is more than that reasonably needed for delegating to y . In this case, x is too prudent and have searched for too many evidences and information. Since also knowledge has costs and utility, in this case the cost of the additional knowledge about y exceeds its utility: x already has enough evidence to delegate. Only in this case the well-grounded trust in y is 'too much'. But notice that we cannot call it 'over-confidence'. In sum, there are three cases of 'too much trust':

- More positive trust in y than necessary for delegating. It is not true that 'I trust y too much' but is the case that I need too much security and information.
- I have more trust in y than he deserves; part of my evaluations and expectations are faithful and unwarranted; I do not see the actual risk. This is a case of *over-confidence*. This is dangerous and irrational trust.
- My evaluation of y is correct but I'm too risk prone; I accept too much ignorance and uncertainty, or I bet too much on a low probability. This is another case of *over-confidence*, and of dangerous and irrational trust.

Which are the consequences of over-confidence in delegation?

- Delegating to an unreliable or incompetent y ;
- Lack of control on y (y does not provide his service, or provide a bad service, etc.);
- Too 'open' delegation: unchecked misunderstandings, y 's inability to plan or to chose, etc.

Which are on the contrary the consequences of insufficient confidence, of an excess of diffidence in delegation?

- We do not delegate and rely on good potential partners; we miss good opportunities; there is a reduction of exchanges and cooperation;
- We search and wait for too many evidences and proofs;
- We make too many controls, loosing time and resources and creating interferences and conflicts;
- We specify too much the task/role without exploiting y 's competence, intelligence, or local information; we create too many rules and norms that interfere with a flexible and opportunistic solution.

So, some diffidence, some lack of trust, prudence and the awareness of being ignorant are obviously useful; but also trusting it is. Which is the right ratio between trust and diffidence? Which is the right degree of trust?

- The right level of positive trust in y (esteem) is when the marginal utility of the additional evidence on y (its contribution for a rational decision) seems inferior to the cost for acquiring it (including time).
- The right degree of trust for delegating (betting) is when the risk that we accept in case of failure is inferior to the expected subjective utility in case of success (the equation -as we saw- is more complex since we have also to take into account alternative possible delegations or actions).

8. Conclusions

We provided a definition of trust as a mental state and presented its *mental ingredients* relative both to the competence of y and to its predictability and x 's faithfulness. *Principled trust requires BDI-like agents*. On the one side, for modelling trust some sort of BDI agents is needed; on the other side, social interaction among BDI-like agents must be based on trust, as a coherent and justified pattern of mental ingredients supporting the intention of delegating and collaborating. We didn't take into consideration other forms of trust (about tools, objects, and natural processes; implicit and procedural forms of trust; affective trust just based on some emotional appraisal - [Tha]. We focused on a cognitive approach to trust for several reasons: its closeness with AI modelling of mind, and our theory of delegation/adoption; its non-reducibility to a simple probability or risk index, and its typical pattern of grounded and connected evaluations and expectations; the link between beliefs and decisions, and the two facets or rationality; the provision

of a basis for a theory of influence and argumentation about trusting and delegating; its analytical power about how increasing trust changing certain beliefs, or how trust is a dynamic process.

We have shown how trust is the mental background of delegation, and their relationship. How and why trust is a bet, and implies risks, has been derived from its reference to a goal, from the action of delegating, and from the uncertainty of trust-beliefs. Both the very basic forms of weak delegation and the more complex forms of social trust, based on morality and reputation, have been analysed. We have discussed some problematic game-theoretical definition, and explained that deep social trust is about prevailing motives in y , and about renouncing to control. Finally we presented a principled quantification of the degree of trust, derived from its cognitive ingredients. The *degree of trust* has been used to model the decision to delegate or not to delegate.

The paper is intended to contribute both to the conceptual analysis and to the practical use of trust in social theory and MAS. The research is based both on the MAS, and on the sociological and socio-psychological literature, although in this paper the discussion of the socio-psychological aspects has been limited. We did not analyse the affective aspects of trust, and we also put aside trust in beliefs and in knowledge sources. Especially the second topic is strongly related to the present contribution, since paradoxically our trust in y is based on our trust in our beliefs about y , which is based on our trust in the *sources* (often social) [Dem] of those beliefs.

References

- [Bon] Bonniver Tuomela, M., (1999), A general account of trust, Tecnical Report, University of Helsinki.
- [Cas1] Castelfranchi, C., Falcone, R., Delegation Conflicts, in M. Boman & W. Van de Velde (eds.) Multi-Agent Rationality, Lecture Notes in Artificial Intelligence, 1237. Springer-Verlag pg.234-254, 1997.
- [Cas2] Castelfranchi, C., Falcone, R., (1998) Towards a Theory of Delegation for Agent-based Systems, *Robotics and Autonomous Systems*, Special issue on Multi-Agent Rationality, Elsevier Editor, Vol 24, Nos 3-4, , pp.141-157.
- [Cas3] C. Castelfranchi, Social Power: a missed point in DAI, MA and HCI. In Decentralized AI. Y. Demazeau & J.P.Mueller (eds) (Elsevier, Amsterdam 1991) 49-62.
- [Cas4] Castelfranchi, C., Conte, R., Limits of economic and strategic Rationality for Agents and M-A Systems. *Robotics and Autonomous Systems*, Special issue on Multi-Agent Rationality, Elsevier Editor, Vol 24, Nos 3-4, , pp.127-139.
- [Cas5] Castelfranchi, C., Commitment: from intentions to groups and organizations. In *Proceedings of ICMAS'96*, S.Francisco, June 1996, AAAI-MIT Press.
- [Cas6] Castelfranchi C., Falcone R., (1998) Principles of trust for MAS: cognitive anatomy, social importance, and quantification, *Proceedings of the International Conference of Multi-Agent Systems (ICMAS'98)*, pp. 72-79, Paris, July.
- [Cas7] Castelfranchi C., (1998), Modelling Social Action for AI Agents, *Artificial Intelligence*, 103, pp. 157-182.
- [Cas8] Castelfranchi C., (1996), Practical Permission, *Workshop on Practical Reasoning and Rationality*, Manchester.
- [Cas9] Castelfranchi C., (1997), Individual Social Action. In G. Holmstrom-Hintikka and R. Tuomela (eds), *Contemporary Theory of action*. Vol II, 163-92. Dordrecht, Kluwer.
- [Cas10] Castelfranchi, C., Commitment: from intentions to groups and organizations. In *Proceedings of ICMAS'95*, S.Francisco, June 1996, AAAI-MIT Press.
- [Cas11] Castelfranchi, C., Conte, R., Paolucci, M., (1997), Normative compliance and the costs of transgression. In Conte, R. and Chattoe, E. (eds), *Evolving societies: the computer simulation of social systems*.
- [CfP] Call for Papers, 1998 *Autonomous Agents '98 Workshop on "Deception, Fraud and Trust in Agent Societes"*, Minneapolis/St Paul, USA, May 9.
- [Con1] Conte, R., Castelfranchi, C., Cognitive and Social Action, (Section 10). London, UCL Press, 1995.
- [Con2] Conte, R., Castelfranchi, C., Dignum, F., Autonomous Norms Acceptance. In J. Mueller, M. Singh, A.S. Rao (eds) *Intelligent Agents V*, Berlin, Springer, 1998, 99-113.
- [Cra] Crabtree B., Wiegand M., Davies J., Building Practical Agent-based Systems, PAAM Tutorial, London, 1996.
- [Das] P. Dasgupta. Trust as a commodity. In D. Gambetta, editor, *Trust*. Chapter 4, pages 49-72. Basil Blackwell, Oxford, 1990.
- [Dem] R. Demolombe, Formalizing the Reliability of Agent's Information, in *Proceedings of the 4th ModelAge Workshop on "Formal Models of Agents"*, 1997.
- [Deu1] M. Deutsch, *The Resolution of Conflict*. Yale University Press, New Haven and London, 1973.
- [Fal] Falcone R., A delegation based theory of agents in organizations, *Mathematical Modelling and Scientific Computing*, Vol. 8, 1997 (ISSN 1067-0688).
- [Gam] D. Gambetta, editor. *Trust*. Basil Blackwell, Oxford, 1990.

- [Hun] Hunt, D. P. and Hassmen, P. *What it means to know something*. Reports from the Department of Psychology, Stockholm University, N. 835, 1997.
- [Jen] Jennings, N.R. (1993). Commitments and conventions: The foundation of coordination in multi-agent systems. *The Knowledge Engineering Review*, 3, 223-50.
- [Lie] Lieberman
- [Lin] B. van Linder, Modal Logics for Rational Agents, PhD thesis, Department of Computing Science, University of Utrecht, 1996.
- [Kap] Kaplan, S. and Garrik, J. On the quantitative definition of risk. In *Risk Analysis*, Vol.1, 1.
- [Luh] N. Luhmann, Familiarity, confidence, trust: Problems and alternatives. In Diego Gambetta, editor, *Trust*. Chapter 6, pages 94-107. Basil Blackwell, Oxford, 1990.
- [Mar] S. Marsh, Formalising Trust as a Computational Concept, PhD thesis, Department of Computing Science, University of Stirling, 1994.
- [Mey] J.J. Ch. Meyer, W. van der Hoek. A modal logic for nonmonotonic 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.
- [Mic] Miceli, M., Cesta, A., Rizzo, P., Distributed Artificial Intelligence from a Socio-Cognitive Standpoint: Looking at Reasons for Interaction. *Artificial Intelligence and Society*, 1995, 9:287-320.
- [Pea] Pears, H. (1971) *What is knowledge ?*. N.Y., Harper and Row.
- [Rau] Raub W, Weesie J., Reputation and Efficiency in Social Interactions: an Example of Network Effects. *American Journal of Sociology* **96**: 626-654, 1990.
- [Sic] Sichman, J, R. Conte, C. Castelfranchi, Y. Demazeau. A social reasoning mechanism based on dependence networks. In *Proceedings of the 11th ECAI*, 1994.
- [Sin] Singh, M., (1999), An ontology for commitments in multiagent systems, *Artificial Intelligence and Law*, Special Issue on "Agents and Norms", Conte, R., Falcone, R., and Sartor, G. (eds), Vol.7, Issue 1, March 1999, pp.97-113.
- [Sni] C. Snijders and G. Keren, Determinants of Trust, *Proceedings of the workshop in honor of Amnon Rapoport*, University of North Carolina at Chapel Hill, USA, 6-7 August, 1996.
- [Tha] Thagard, P., (1998), *Emotional Coherence: Trust, Empathy, Nationalism, Weakness of Will, Beauty, Humor, and Cognitive Therapy*, Technical Report, University of Waterloo.
- [Tuo] Tuomela, R., (1995), *The importance of us: a philosophical study of basic social notions*, Stanford University Press.
- [Wil] Williamson, O. E., (1985), *The Economic Institutions of Capitalism*, The Free press, New York, 1985.