Trust Is Much More than Subjective Probability: Mental Components and Sources of Trust

Cristiano Castelfranchi and Rino Falcone
National Research Council - Institute of Psychology
Unit of "AI, Cognitive Modeling and Interaction"
Roma - Italy
e-mail: falcone@ip.rm.cnr.it, cris@pscs2.irmkant.rm.cnr.it

Abstract

In this paper we claim the importance of a cognitive view of trust (its articulate, analytic and founded view), in contrast with a mere quantitative and opaque view of trust supported by Economics and Game Theory (GT). We argue in favour of a cognitive view of trust as a complex structure of beliefs and goals, implying that the trustor must have a "theory of the mind" of the trustee. Such a structure of beliefs determines a "degree of trust" and an estimation of risk, and then a decision to rely or not on the other, which is also based on a personal threshold of risk acceptance/avoidance. Finally, we also explain rational and irrational components and uses of trust.

1. Introduction

Is it trust simply reducible to "subjective probability" ?

This is in fact the dominant tradition in Economics, Game Theory, part of Sociology [Gam-88; Col-94], and now in Artificial Intelligence and Electronic Commerce [Bra-99]. We argue in favour of *a cognitive view of trust* as a complex structure of beliefs and goals (in particular causal attributions, evaluations and expectations), implying that the trustor must have a "theory of the mind" of the trustee (possibly including personality, shared values, morality, goodwill, etc.) [Cas-98, Cas-99].

Such a structure of beliefs determines a "degree of trust" and an estimation of risk, and then a decision to rely or not on the other, which is also based on a personal threshold of risk acceptance/avoidance.

In this paper we use this mental model of trust for two claims.

On the one side, we claim that there are several *sources* of the beliefs on which the trust is based, and that the basis and the dynamics of trust cannot be reduced to reinforcement learning or probability updating on the basis of personal experience and personal interactions (although this is an important source) [Jon-99, Bis-99]. Trust beliefs come also from other sources: from observations, reasoning, social stereotypes, communication, spreading of reputation, signs [Bac-99], etc. We provide a model of these sources, and of the relationship between trust in information sources and social trust in delegation.

On the other side, we argue against probability reduction and "eliminativism". We agree with Williamson [Wil-85] that one can/should eliminate the redundant, vague, and humanistic notion of 'trust' *if* it simply covers the use of subjective probability in decisions. But we strongly argue against both this reduction and the consequent elimination. Trust cannot be reduced to a simple and opaque index of probability because agents' decisions and behaviours depend on the specific, qualitative evaluations and mental components. For example, internal or external attribution of risk/success, or a differential evaluation of trustee's competence Vs willingness, make very different predictions both about trustor's decisions and possible interventions and cautions.

All these mental and dispositional components of trust are quite relevant also in electronic commerce where, for example, we have to distinguish between different sources and reasons for caution or distrust. In particular, a rich cognitive analysis of trust is coherent with a cognitive view of Agents (ex. Belief Desire Intention approach [Bra-87, Had-96]), an important role given to norms, expectations, roles, etc.; a socially situated view of agents in communities and institutions [Gan-99].

2. Limits of the Strategic Tradition on Trust

Doubtless the most important tradition of studies on trust is the "strategic" tradition, which builds upon the rational

decision and Game Theories to provide us a theory of trust in conflict resolution, diplomacy, etc. and also in commerce, agency, and in general in economics. Let us try to discuss two positions, one strongly relating trust and cooperation in Prisoner's Dilemma (PD) situations, the other more extreme which denies the utility of the notion of trust in favour of subjective probability or risk.

2.1. Is Trust only a Prisoner's Dilemma?

Deutsch's definition [Deu-58] of trust in terms of expectations well represents strategic tradition:

An individual may be said to have trust in the occurrence of an event if he expects its occurrence and his expectations lead to behaviour which he perceives to have greater negative consequences if the expectation is not confirmed than positive motivational experiences if it is confirmed.

Although we agree about the importance of *expectation* in trust, let us notice that this definition completely ignores the *evaluation component*, which makes such an expectation reason-based 1.

However, the most important aspect of this definition is the arbitrary restriction of trust to situations where the risks are greater than the utility. This does not correspond to common sense and natural language and it is not justified as a technical terminological decision by some heuristic advantage. In fact -as remarked by Coleman [Col-94]- several important examples and natural situations of trust would be excluded without any advantage. What is really important in Deutsch's analysis is the notion of *vulnerability*, the fact that the trustor is (and feels itself) exposed to danger; the idea that in trust necessarily there are risks and uncertainty. More than this: it is true that *the act itself of trusting and relying on exposes to risks*. However, this is not due to a PD-like situation where to defeat -when the other is cooperating- pays more than cooperating. It is much more general: by deciding of trusting the other agent, I expose myself to risks because I decide to bet on the other (while if I do not bet on it, if I do not delegate, I will not risk).

We do not believe that Deutsch had in mind the general fact that the failure of whatever action or plan (including a delegation) results not only in the unfulfilled goal, in the unrealised expected utility, but also in some loss (the invested resources and missed opportunities). If we consider the negative utility in case of failure as equivalent to the expected benefit or utility in case of achievement 2 , given losses (costs), it is always true (in any decision and in any action) that the negative outcome of failure is greater than the positive outcome of success. Deutsch does not want to be so general; he precisely intends to restrict trust to a special class of strategic situations. In fact, there are situations where in case of failure there are additional losses: not simply the invested resources (included time and thinking) are lost, and the motivating goals are unachieved, but some other goal is damaged. Consider for example a failure which implies shame and bad reputation. PD is an example of these situations, since the damage of delegating and fail (*x*'s cooperation and *y*'s defection) is greater than the damage of not delegating at all (*x*'s non cooperation).

Of course *the greater the damage in case of failure the greater the risk.* But *trust applies to any risky situation*, i.e. to any uncertain decision and plan, not only in the very unbalanced situation with additional risks. For sure, *the greater the perceived risk, the greater the needed trust* in order to trust, but also decisions with small risks require some trust. Trust is involved in usual, every day decisions. Deutsch wants trust be special and outside rational decisions: following rational decision criteria one shouldn't rely on, shouldn't bet on that course of events, as for the so called "cooperative" move in the Prisoner Dilemma; if one does it is just because of trust (see later).

There is a correct and important intuition in Deutsch that should be accounted for (the idea that the greater the perceived risk, the greater the needed trust to trust); but why to make this relationship discontinuous (only if the risk is greater than the utility, there is trust)? Trust (or better the degree of trust) is a continuous, an agent can trust an event also if the resulting utility is greater than the connected risk (undoubtedly, the needed trust is not so big as in the opposite case). Only the decision to trust (or better to delegate a task, see [Cas-98b]) is discrete: either trust is sufficient or it is not, either I (decide to) trust or not.

Moreover, trust can be irrational but it is not necessarily so. Notice that if to trust would be always non rational, when and how would trust be "insufficient" to rely on? About this question, in our model (see later) there are both a ratio between utility and risk (that makes the decision rational or not), and an idiosyncratic factor of risk avoidance or acceptance that makes the degree of trust individually sufficient or not to trust [Cas-99].

Analogous view of trust we find in the conclusion of Gambetta's book [Gam-90] and in [Bac, Bac-99]: "In general, we say that a person 'trusts someone to do α' if she acts on the expectation that he will do α when two conditions obtain: both know that if he fails to do α she would have done better to act otherwise, and her acting in the way she does gives him a selfish reason not to do α."

Also in this definition we recognise the 'Prisoner's Dilemma syndrome' that gives an artificially limited and quite pessimistic view of social interaction. In fact, by trusting the other makes herself 'vulnerable'; in other terms, she gives to the other the *opportunity* to damage her. As we

just said this is true, but not necessarily she gives him a *motive*, a reason for damaging her (on the contrary, in some cases to trust someone represents an opportunity for the trustee to show his competencies, abilities, willingness, etc.).

Not necessarily there is trust only if trusting the other makes convenient for him to disappoint trustor's expectation. Perhaps trustor's trusting him gives him (the trustee) a reason and a motive for not disappointing trustor's expectation; perhaps trustor's delegation makes the expected behaviour of trustee convenient for the trustee himself, it could create an opportunity for cooperation on a common goal. Trust continues to be trust independently of making convenient or not for the trustee to disappoint the trustor. Of course, there could be always risks and uncertainty, but not necessarily *conflict* in the trustee between selfish interest and broader or collective interests. If this were true there were no trust in strict cooperation based on common goal, mutual dependence, common interest to cooperate, and a joint plan to achieve the common goal [Con-95]. While on the contrary there is trust in any joint plan, since the success of the trustor depends on the action of the trustee, and vice versa, and the agents are relying on each other.

The strategic view of trust is not general; it is based on an arbitrary and unproductive restriction. It is interested only in those situations where additional (moral or contractual) motivations, additional external incentives are needed. While in several cases intentions, intention declarations, esteem, goodwill are enough.

The strategic view also exposes itself to a serious attack, aimed at the elimination of the notion of trust. We will see this (¤2.2.) and how a cognitive analysis of trust resists to it (¤3.)

2.2. Against eliminativism: in defence of (a cognitive theory of) *trust*

The traditional arrogance of economics and its attempt to colonialise with its robust apparatus social theory (political theory, theory of law, theory of organisations, theory of family, etc.[1]) coherently arrives - on the field of 'trust' - to a 'collision' [Wil-85] with the sociological view.

The claim is that the notion of 'trust' when applied in the economic and organisational domain or, in general, in strategic interactions is just a common sense, empty term without any scientific added value[2] ; and that the traditional notions provided by transaction cost economics are more 'parsimonious' and completely sufficient for accounting for and explaining all those situations where lay people (and sociologists) use the term 'trust' (except for very special and few personal and affective relationships[3]). The term trust is just for suggestion, for making the theory more 'user-friendly' and less cynic. It is just 'rhetoric' when applied to commerce[4] but does not explain nothing about its nature which is and must be merely 'calculative' and 'cynic'[5] .

On the one side, we should say that Williamson is pretty right: if trust is simply subjective probability, or if what is useful and interesting in trust is simply the (implicit) subjective probability (like in Gambetta's definition [Gam-88] -see note 6- and in the game-theoretic and rational decision use of trust), then the notion of trust is redundant, useless and even misleading. On the other side, the fact is that trust is not simply this, and -more important- what of the notion of trust is useful in the theory of social interactions is not only subjective probability.

---

1 In his section on 'Economics and the Contiguous Disciplines' (p.251) [Wil-85] Williamson himself gives example of this in law, political science, in sociology.

2 'There is no obvious value added by describing a decision to accept a risk (...) as one of trust' [Wil-85, p.265]. 'Reference to trust adds nothing' [Wil-85, p.265].

3 '(...) trust, if obtains at all, is reserved for very special relations between family, friends, and lovers' [Wil-85, p.273].

4 'I argue that it is *redundant* at best and can be *misleading* to use the term "trust" to describe commercial exchange (...) Calculative trust is a contradiction in terms' [Wil-85, p. 256].
'(...) the *rhetoric* of exchange often employs the language of promises, trust, favors, and cooperativness. That is understandable, in that the *artful* use of language can produce deals that would be scuttled by abrasive calculativness. If however the basic deal is shaped by objective factors, then calculativness (credibility, hazard, safeguards, net benefits) is where the crucial action resides.' [Wil-85, p. 260].
'If calculative relations are best described in calculative terms, then the diffuse terms, of which trust in one, that have mixed meanings should be avoided when possible.' [Wil-85, p.261] And this does not apply only to the economic examples but also to the apparent exception of 'the assault girl (...) I contend is not properly described as a condition of trust either' [Wil-85, p.261]. This example that is 'mainly explained by bounded rationality - the risk was taken because the girl did not get the calculus right or because she was not cleaver enough to devise a contrived but polite refusal on the spot - is not illuminated by appealing to trust'. [Wil-85, p. 267].

5 'Not only is "calculated trust" a contradiction in term, but *user friendly terms*, of which "trust" is one, have an additional cost. The world of commerce is reorganised in favor of the cynics, as against the innocents, when social scientists employ user-friendly language that is not descriptively accurate - since only the innocents are taken in' [Wil-85, p.274].
In other words, "trust" terminology edulcorates and masks the *cynic reality* of commerce. Notice how Williamson is here quite prescriptive and neither normative nor descriptive about the real nature of commerce and of the mental attitudes of real actors in it.

Not only Williamson is assuming more a prescriptive than a scientific descriptive or explanatory attitude, but he is simply wrong in his eliminativistic claims. And he is wrong even about the economic domain, which in fact is and must obviously be socially embedded. Socially embedded does not mean only -as Williamson claims- institutions, norms, culture, etc.; but also means that *the economic actors are fully social actors* and that they act in such a habit also in economic transactions, i.e. with all their motives, ideas, relationships, etc. included the *trust* they have or not in their partners and in the institutions.

The fact that he is unable to see what 'trust' adds to the economic analysis of risk[6] , and that he considers those terms as equivalent, simply shows how he is unable to take into account the interest and the contribution of cognitive theory.

Risk is just about the possible outcome of a choice, about an event and a result; trust is about somebody: it mainly consists of beliefs, evaluations, and expectations about the other actor, his capabilities, self-confidence, willingness, persistence, morality (and in general motivations), goals and beliefs, etc. Trust *in* somebody basically is (or better at least includes and is based on) a rich and complex theory of him and of his mind. Conversely distrust or mistrust is not simply a pessimistic esteem of probability: it is diffidence, suspect, negative evaluations *relative to* somebody.

For his traditional economic perspective all this is both superfluous and naive (non-scientific, rhetoric): common-sense notions. He does not want to admit the insufficiency of the economic theoretical apparatus and the opportunity of its cognitive completion.

But he is wrong -even within the economic domain- not only for the growing interest in economics for more realistic and psychologically based model of the economic actor, but because mental representations of the economic agents and their social images are -for example- precisely the topic of marketing and advertising (that we suppose have something to do with commerce).

His claim about parsimony, sufficiency, and the absence of 'added value' is quite strange from a methodological point of view. In fact, a given description of *X* is parsimonious and adequate, sufficient or insufficient, only relative to given purposes the description is for. He should at most claim that for the purposes of the economic analysis the transaction cost framework is necessary and sufficient and that 'trust' does not add anything relevant *for the economic perspective* (it is just a cosmetic bla-bla). But this is not his claim. His claim pretends to be general, to provide *the* correct and sufficient interpretation of the situations. In fact it borrows the examples he analyses from sociology and he does not concede that analysing those situations in terms of trust would add something relevant at least for the social or cognitive theory! (this is why we used the term 'arrogance' about economics).

On the contrary, we claim that analysing trust and analysing those situations in terms of trust is absolutely necessary for modelling and explaining them from a psychological, anthropological or sociological scientific perspective. We claim that the richness of the mental ingredients of trust cannot and should not be compressed simply in the subjective probability estimated by the actor for his decision. But why do we need an explicit account of the mental ingredients of trust (beliefs, evaluations, expectations, goals, motivations, model of the other), i.e. of the *mental background* of reliance and of 'probability' and 'risk' components?

- First, because otherwise we will neither be able to explain or to predict the agent's risk perception and decision. Subjective probability is not a magic and arbitrary number; it is the consequence of the actor beliefs and theories about the world and the other agents.
- Second, because without an explicit theory of the cognitive bases of trust any theory of persuasion/dissuasion, influence, signs and images for trust, deception, reputation, etc. is not 'parsimonious' but is simply empty.

Let's supposed that the girl under risk of assault is Mr. Brown's daughter D and that Mr Brown is an anxious father, and that he has also a son from the same school of that guy G accompanying the girl. Will he ask for his son "Which is the probability that G assault your sister D?" We do not think so. He will ask for his son what he knows *about* G, if he has evaluation/information about G's education, his character, his morality, etc. And this not for rhetoric or for

---

[6] Section 2. starts with 'My purpose in this and the next sections is to examine the (...) "elusive notion of trust". That will be facilitated by examining a series of examples in which *the terms trust and risk are used interchangeably - which has come to be standard practice in the social science literature* - (...)'. The title of section 2.1 is in fact 'Trust as Risk'.
Williamson is right in the last claim. This emptying of the notion of trust is not only his own aim, it is quite traditional in sociological and game-theoretic approaches. For example in the conclusions of his famous book [Gam-88] Gambetta says: '.. *When we say we trust someone or that someone is trustworthy, we implicitly mean that the probability* that he will perform an action that is beneficial or at least not detrimental to us is high enough for us to consider engaging in some form of cooperation with him' [Gam-88, p.217]. What is dramatically not clear in this view is what "trust" does *explicitely* mean! In fact the expression cited by Williamson (the 'elusive notion of trust') is from Gambetta.
His objective is the elimination of the notion of trust from economic and social theory (it can perhaps survive in social psychology of interpersonal relationships). 'The recent tendency for sociologists /the attack is mainly to Coleman and to Gambetta/ and economists alike to use the term "trust" and "risk" interchangeably is, on the arguments advanced here, ill-advised'.

using a more friendly notion. This is because he searches for some specific and contentfull information able to found his prediction/expectation about risk. Coleman too stresses the importance of information, but he is not able to derive from this the right theoretical consequences: a view of trust also in terms of justified cognitive evaluations and expectations. In his theory one cannot explain or predict which information is pertinent and why. For example, why the artistic talent of G or the colour of his car are irrelevant.

Now what is the relation between this information *about* G and the estimated risk or probability? Is Williamson's theory able to explain and predict this relation? In his framework subjective probability and risk is an *unprincipled and ungrounded notion*. What the notion of trust (its cognitive analysis) adds to this framework is precisely the explicit theory of the ground and (more or less rational) support of the actor's expectation, i.e. the theory of a specific set of beliefs and evaluations *about* G (the trustee) and about the environmental circumstances, and possibly even of the emotional appraisal of both, such that an actor makes a given estimation of probability of success or failure, and decides whether relying and depending on G or not.

Analogously, what to do in Williamson's framework for acting upon the probability (either objective or subjective)? Is there any rational and principled way? He can just to touch wood or make exorcism to try to modify this magic number of the predicted probability. Why and how should for example information about 'honesty' change my perceived risk and my expected probability of an action of G? Why and how should for example training, friendship, promises, a contract, norms[7] , or control, and so on, affect (increase) the probability of a given successful action and my estimation of it? It remains unexplained.

In the economic framework, first we can only account for a part of these factors, second this account is quite incomplete and unsatisfactory.

We can account only for those factors that affect the rewards of the actor and then the probability that he will prefer one action to another. Honour, norms, friendship, promises, etc. must be translate into positive or negative 'incentives' on choice (for ex. to cooperate Vs to defeat). This account is very reductive. In fact, we do not understand in the theory how and why a belief (information) about the existence of a given norm or control, or of a given treat, can generate a goal in G's mind and eventually change his preferences. Notice on the contrary that our predictions and our actions of influencing are precisely based on a 'theory' of this, on a 'theory' of G's mind and mental processes beyond and underlying 'calculation'. Calculation is not only institutionally but also *cognitively embedded* and justified!

Other important aspects seem completely out of the theory. For example the ability and self-confidence of G, and the actions for improving them (for example a training) and for modifying the probability of success, or the action for acquiring information about this and increase the subjective estimated probability.

Trust is also this: beliefs about G's competence and level of ability, and his self-confidence. And this is a very important basis for the prediction and esteem of the probability of success or the risk of failure.

3. A Cognitive Analysis of Trust

Let us introduce briefly our cognitive analysis of trust (for a more complete presentation see [Cas-98a, Cas-98b, Cas-98c, Cas-99]). In our model we specify which beliefs and which goals characterise *x*'s trust in another agent *y*.

3.1. Beliefs on which Trust is based

Only a cognitive agent can "trust" another agent. We mean: only an agent endowed with goals and beliefs.

First, one trusts another only relatively to a goal, i.e. for something s/he wants to achieve, that s/he desires. If *x* does not have goals, she cannot really decide, nor care about something (welfare): she cannot subjectively "trust" somebody.

Second, trust itself *consists of* beliefs. Trust basically is *a mental state*, a complex mental *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 (trustor), it is a cognitive agent endowed with* internal explicit goals and beliefs;
- *y* is the agent or entity which is trusted *(trustee)*; *y* is not necessarily a cognitive agent (in this paper, however, we will consider only cognitive agents). So
- *x* trusts *y* "about" g/α (where g is a specific world state, and α is an action that produces that world state g) 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. *Trust is the mental counter-part of delegation.*

We summarize the main beliefs in our model (their relationships are better explained in [Cas-99]:

1. "Competence" Belief: a *positive evaluation* of *y* is necessary, *x* should believe that *y* is useful for this goal of

[7] How and why 'regulation can serve to *infuse* trading *confidence* (i.e. trust!!) into otherwise problematic trading relations' as Williamson reminds by citing Goldberg and Zucker (p. 268).

hers, that *y* can produce/provide the expected result, that *y* can play such a role in her plan/action, that *y* has some function.

2. "Disposition" Belief: Moreover, *x* should believe that *y* is not only able to perform that action/task, but *y* will actually do what *x* needs. With cognitive agents this will be a belief with respect to their *willingness*: this make them predictable.

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

4. Fulfilment Belief: *x* believes that *g* will be achieved (thanks to *y* in this case)[8] . This is the "trust that" g.

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

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

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

We can say that trust is a set of mental attitudes characterizing the mind of a "delegating" agent, who 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 versus external attribution of Trust

We should also distinguish between trust 'in' someone or something that has to act and produce a given performance thanks to its *internal* characteristics, and the global trust in the global event or process and its result which is also affected by external factors like opportunities and interferences.

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*, and *disposition*. Evaluation of *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***.**

This distinction between internal versus external attribution is important for several reasons:

- To better capture the meaning of trust in several common sense and social science uses.
- To understand the precise role of that nucleus of trust that we could describe in terms of "unharmfulness", sense of safety, perception of goodwill.
- 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 rather to (either implicitly or explicitly) imply, the *subjective probability* of the successful performance of a given behaviour α, and it is on the basis of this subjective perception/evaluation of risk and opportunity that the agent decides to rely or not, to bet or not on *y*. However, the probability index is based on, derived 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 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 is this decomposition 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 also because:

a) the agent's trusting/delegating decision might be different with the same global probability or risk, depending on its composition;

b) trust composition (internal versus external) produces completely different intervention strategies: manipulating the external variables (circumstances, infrastructures) is completely different from manipulating internal parameters.

Let us consider the first point. There might be different heuristics or different personalities with a 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

[8] The trust that g does not necessarily requires the trust in y. x might ignore which are the causal factors producing or maintaining g true in the world, nevertheless x 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.

problem is that - given the same global expectation - one agent might decide to trust/rely in one case but not in the other, or vice versa!

As for point (b), the strategies to establish or increment trust are very different depending on the external or internal attribution of your diagnosis of lack of trust. If there are adverse environmental or situational conditions your intervention will be in establishing protection conditions and guarantees, 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..

We should also consider the reciprocal influence between external and internal factors. When x trusts the internal powers of y, x also trusts y's abilities to create positive opportunities for success, to perceive and react to the external problems. Vice versa, when x trusts the environment opportunities, this evaluation could change the trust in 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) are aspects of the external trust. Is it important to stress that:

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

Vice versa, when I strongly trust y, i.e. his abilities, willingness and faithfulness, I can accept a less safe and reliable environment (with less external monitoring and authority). We account for this 'complementariety' [Gan-99] between the internal and the external components of trust in y for g in given circumstances and a given environment.

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

3.3. Degrees of Trust

The idea that trust is scalable is common (in common sense, in social sciences, in AI). 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. 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. More precisely the latter are based on the former.

In our model we ground the degree of trust of *x* in *y*, in the cognitive components of *x*'s mental state of trust. More precisely, *the degree of trust is a function of the subjective certainty of the pertinent beliefs*. We use the degree of trust to formalise a rational basis for the decision of relying and betting on *y*. Also we claim that the "quantitative" aspect of another basic ingredient is relevant: *the value or importance or utility of the goal g*. In sum,

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

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

Let's call the degree of trust of *X* in *Y* about τ: $\mathbf{DoT_{XY\tau}}(0 \leq DoT_{XY\tau} \leq 1)$. 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.

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

where:

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

$DoC_X[WillDo_Y(\alpha,g)] = DoC_X[Intend_Y(\alpha,g)] * DoC_X[Persist_Y(\alpha,g)]$
(given that Y is a *cognitive agent*)

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

3.4. 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 represent the "rational" (reasons based, see 3.6.) part of the trust in y. But those beliefs can also be not really warranted, not based on evidences, quite irrational, faithful. We call this part of the trust in y: "faith".

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

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 degree of trust: its sum should be greater than discouraging factors. We are interested here in the additional fact that these (grounded or ungrounded) positive expectations are not enough for explaining the *decision/act* of trusting. In fact, another aspect is necessarily involved in this decision. 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: there always remains some uncertainty (ignorance) and some probability of failure, and the agent must accept this and run a risk.

Thus, a fundamental component of our decision to trust y, is our acceptance and felt exposition to risk. Risk is represented in the quantification of the degree of trust 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.

In our model [Cas-99] we introduce not only a "rational" degree of trust but also a parameter able to evaluate the risk factor. In fact, in several situations and contexts, not just for the human decision makers but -we think- also for good artificial decision makers, it should be important to consider the absolute values of some parameter independently from the values of the others. This fact suggests the introduction of 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 is independently from the probability of the failure (even if it is very low) and from the possible payoff (even if it is very high). In other words, that danger might seem to the agent an intolerable risk (for example, in our model we introduce an 'acceptable damage' threshold).

3.5. Rational trust

In our view trust can be rational and can support rational decisions. Trust as attitude 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.

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, it is not well supported. Either the faith[9] component (unwarranted expectations) or the risk acceptance (blind trust) are too high[10] .

In a sense Deutsch is interested only in one facet of trust, in one form and meaning of it: what we called the dark-side of trust or blind trust.

3.5.1. When trust is too few or too much: 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 the trustor accepts too much risk or too much ignorance, or is not accurate in her evaluations. Notice that there cannot be too much positive trust, esteem of the trustee. It could be not well grounded: the actual risk is greater than the subjective one. Positive evaluation on the trustee (trust in him) can be too much only in the sense that it is more than that reasonably needed for delegating to him. In this case, the trustor is too prudent and has searched for too many evidences and information. Since also knowledge has costs and utility, in this case the cost of the additional knowledge about the trustee could exceed its utility: the trustor already has enough evidence to delegate. Only in this case the well-grounded trust in the trustee 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 the trustee than necessary for delegating. It is not true that 'the trustor trusts the

---

[9] Non-rational blind trust is close to faith. Faith is more that trust without evidences, it is trust whitouht the need for and the search for evidences.

[10] 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 cannot be based on some understanding of the process or some model of it, but just based on repeated experiences and assotiations.

trustee too much' but is the case that she needs too much security and information.

- The trustor has more trust in the trustee than he deserves; part of my evaluations and expectations are unwarranted; I do not see the actual risk. This is a case of *over-confidence*. This is dangerous and irrational trust.
- Trustor's evaluation of the trustee is correct but she is too risk prone; she accepts too much ignorance and uncertainty, or she bets 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 trustee;
- Lack of control on the trustee (he does not provide his service, or provide a bad service, etc.);
- Too *open delegation* [Cas-98c]: in other words, a delegation that permits (obligates) the trustee to make chooses, plans, etc., and he is unable to realize such a kind of actions.

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 trustee'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 the trustee (esteem) is when the marginal utility of the additional evidence on him (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 in [Cas-99]- is more complex since we have also to take into account alternative possible delegations or actions).

4. Concluding remarks

In sum, if trust is nothing more than subjective probability, as it is in the Game-theoretic tradition and in Coleman, we believe that Williamson's eliminativistic proposal is correct. However, we have argued that the reduction of trust to a simple number, quantity, and specifically to probability is highly unsatisfactory. Trust is a rich and complex mental attitude of x towards y as for a given action and goal. This attitude basically consists of evaluations of y and of the situation, and of expectations about y's mind, behaviour and possible results. These evaluations and expectations are simply some kinds of beliefs, and are based and justified on other beliefs about facts and information sources. Different aspects of this mental representation of y and of the situation are very relevant and make different predictions: for example, the difference between internal and external; attribution; or the distinction between y's competence and willingness. The reduction of these articulated aspects of trust to a probability measure, reduces the explanatory and predictive power of the theory. For example, one is neither able to predict which information about y is pertinent for modifying the trust in y; nor to explain why/how certain circumstances (like the existence of friendship or of enforceable norms) make y more predictable; nor eventually to specify which aspects or part of the trust are irrational.

5. References

[Bac-99] Bacharach M. and Gambetta D., Trust as Type Interpretation, in Castelfranchi C. and Tan (eds), Trust and Deception in Virtual Societies, Kluwer Publisher, (in press).

[Bac] M. Bacharach and D. Gambetta,*Trust in Signs*, in Karen Cook (ed.) Trust and Social Structure, New York: Russel Sage Foundation, forthcoming.

[Bis-99] A. Biswas, S. Sen, S. Debnath, (1999), Limiting Deception in Social Agent-Group, *Autonomous Agents '99 Workshop on "Deception, Fraud and Trust in Agent Societes"*, Seattle, USA, May 1, 21-28.

[Bra-87] M.E. Bratman, (1987), *Intentions, Plans and Practical Reason*. Harward University Press: Cambridge, MA.

[Bra-99] S. Brainov and T. Sandholm, (1999), Contracting with uncertain level of trust, Proceedings of the AA'99 Workshop on "Deception, Fraud and Trust in Agent Societies", Seattle, WA, 29-40.

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

[Cas-98b] Castelfranchi C., Falcone R., (1998) Social Trust: cognitive anatomy, social importance, quantification and dynamics, *Autonomous Agents '98 Workshop on "Deception,*

Fraud and Trust in Agent Societes", Minneapolis/St Paul, USA, May 9, pp.35-49.

[Cas-98c] 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.

[Cas-99] Castelfranchi, C., Falcone, R. (1999). The Dynamics of Trust: from Beliefs to Action, *Autonomous Agents '99 Workshop on "Deception, Fraud and Trust in Agent Societes"*, Seattle, USA, May 1, 41-54.

[Col-94] J. S. Coleman, (1994), Foundations of Social Theory, Harvard University Press, MA.

[Con-95] Conte, R., Castelfranchi, C., Cognitive and Social Action, London, UCL Press, 1995.

[Dem-98] R. Demolombe, (1998), To trust information sources: a proposal for a modal logical framework, *Autonomous Agents '98 Workshop on "Deception, Fraud and Trust in Agent Societes"*, Minneapolis, USA, May 9, 9-19.

[Deu-58] M. Deutsch, (1958), Trust and Suspicion *Journal of Conflict Resolution* , Vol. 2 (4) 265-79.

[Gam-90] D. Gambetta, editor. Trust. Basil Blackwell, Oxford, 1990.

[Gam-88] D. Gambetta, (1988), Can we trust trust?, in D. Gambetta (ed.), Trust, Making and Breaking Cooperative Relations, Oxford: Basil Blackwell, 213-237.

[Gan-99] A. Ganzaroli, Y.H. Tan, W. Thoen, (1999), The Social and Institutional Context of Trust in Electronic Commerce, *Autonomous Agents '99 Workshop on "Deception, Fraud and Trust in Agent Societes"*, Seattle, USA, May 1, 65-76.

[Had-96] A. Haddadi, K. Sundermeyer, (1996), Belief-Desire-Intention Agent Architectures, in G.M.P. O'Hare and N.R. Jennings (eds), Foundations of Distributed Artificial Intelligence, Wiley Interscience, pp. 169-186.

[Jon-99] C. Jonker and J. Treur, (1999), Formal Analysis of Models for the Dynamics of Trust based on Experiences, *Autonomous Agents '99 Workshop on "Deception, Fraud and Trust in Agent Societes"*, Seattle, USA, May 1, 81-94.

[Wil-85] Williamson, O.E., (1985), The Economic Institutions of Capitalism, The Free Press, New York.