

# Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification

Cristiano Castelfranchi and Rino Falcone

Division of "AI, Cognitive Modeling and Interaction"  
National Research Council - Institute of Psychology - Roma - Italy  
e-mail: {cris, falcone}@pscs2.irmkant.rm.cnr.it

## 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 most basic forms of non-social trust (reliance on objects and tools) to arrive at the more complex forms of social trust, based on morality and reputation. Finally 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. 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

We will not argue about the absolutely crucial role of trust for the whole human social life and interaction. This would be too obvious (see for ex. [1, 2]). Let's shortly consider exactly the same issue from the point of view of "agents", MAS, and of their foundation. The schema of our argument is quite simple: since

- the notion of agent itself (in AI) implies the notions of delegation, task, and "on the behalf of" [3, 4, 5]; since
- exploitation, exchange, strict cooperation, collaboration and team work, organisations, roles, market, and so on, all the most important forms of sociality and MAS issues are strictly based on some form of delegation of some agent on other agents [6]; and since
- delegation is strictly based on trust;

thus, the notion of trust is crucial in agents' theory and in MAS. Let's shortly address these three claims.

First, although there are many definitions of "agent" in AI, some of which in full disagreement with each other, the majority of them are based on the notions of *task*, of *"behalf of"* or explicitly talk about *delegation*.<sup>1</sup>

Second, in any *contract*, *exchange*, *collaboration*, etc. there is a "client" and a "contractor" [13], and the client is

delegating some task to and relying on the contractor. Organisations are based on *roles* and roles are *delegated* classes of tasks [6]. All this is based on *social commitments*, but they are commitments to do something for the other; they presuppose that the other is relying on us and then delegating us. *Norms* are group prescriptions to do or not to do given actions: they *rely on* the obedience of the addressees and on their behaviour. Clearly enough what we call reliance or delegation (in its weak and strong forms) is a pillar of social cooperation.<sup>2</sup>

Finally, delegation is a M-A 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". We will illustrate this analytically later on. Let's just say here that in order to delegate a task to some agents  $y$  (tool or 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  $\alpha$  or a goal  $g$  to  $y$ , represents precisely my action of entrusting  $y$  for  $\alpha/g$ .

Consider for example the definition of trust provided in the classic book of Gambetta and accepted the great majority of the authors [1]: "*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 to rely on B; and it doesn't explain what is such an evaluation made of and based on: the *subjective probability* includes together too many important parameters and beliefs, which are very relevant in social reasoning.

<sup>1</sup> See [7],[8],[9], [10] and [12]. In particular in [11] J. Ousterhout says: "... The agents need to be able to make decisions that are complex and subtle, and we need to be able to *trust* them enough that we don't have to check up on them constantly". Also in [11] J. White says: "... The two principal meanings of the term agent, which are captured in the adjectives "intelligent" and "mobile", have in common the objective of moving the user beyond interaction with computers to *delegation of responsibility to computers* ...".

<sup>2</sup> Axelrod [14] claimed that trust is not necessary for cooperation, and that cooperation can emerge among selfish and untrustful agents. This is true, but it concerns only "strong delegation" and certain types of cooperation, for example unaware, accidental or functional cooperation, emergent or orchestrated forms, where the participants in the "joint" plan may even ignore each other [15, 16]. The claim is also based on a quite strange notion of "cooperating" which is the so called "cooperative" move into the PD game. But it is not true for higher level cooperation where the participants are aware of each the other, subjectively rely on the share of the others, and usually have an agreement about this, or try to regulate their joint activity by commitments, roles, norms.

## 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

*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 want to achieve, that s/he desires. If I don't potentially have goals, I cannot really decide, nor care about something: I cannot subjectively “trust” somebody.

Second, trust itself *consists of* beliefs. Trust is *a mental state*, a complex *attitude* of an agent  $x$  towards another agent  $y$  about the behaviour/action  $\alpha$  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. It might also be an object or a tool involved in an action of  $x$ , or a natural force or event. The only relevant thing is that: a) it is something able to cause some effect  $g$  through some action  $\alpha$ ; b) this effect is useful for  $x$  (a goal of  $x$ ) and  $x$  is relying on  $y$  for it; so
- $x$  trusts  $y$  about  $g/\alpha$  and for  $g/\alpha$ ;  $x$  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.*

Given this strict relation and the foundational role of delegation (see §1.) 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

In *Delegation*  $x$  needs or likes an action of  $y$  and includes it in her own plan: she relies on  $y$ . She 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 [4].

To do this she has some trust both in  $y$ 's ability and in  $y$ 's predictability, and she should abstain from doing and from delegating to others the same task.

In *weak delegation* there is no bilateral awareness of the delegation, no agreement:  $y$  is not aware of the fact that  $x$  is exploiting her action. One can even “delegate” some task to an object or *tool*, relying on it for some support and result [17; 18]. As an example of weak and passive but already social delegation, which is the simplest form of

social delegation, consider a hunter who is ready to shoot an arrow at a flying bird. In his plan the hunter includes an action of the bird: to continue to fly in the same 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, 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( $x$   $y$   $\tau$ )** to denote a specific *mental state* compound of other more elementary mental attitudes (beliefs, goals, etc.). While we use a predicate **Delegate( $x$   $y$   $\tau$ )** to denote the *action* and the resulting relation between  $x$  and  $y$ .

Delegation necessarily is an *action*, a result of a decision, and it too creates and is a (*social*) *relation* among  $x$ ,  $y$ , and  $\alpha$ . 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).<sup>3</sup>

- 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).

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

- There may be *delegation without trust*:

<sup>3</sup> 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.

these are exceptional cases in which either the delegating agent is not free (*coercive delegation*<sup>4</sup>) 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 I delegate or I do not delegate<sup>5</sup>. Indeed trust has degrees: I trust *y* more or less relatively to  $\alpha$ . And there is a threshold under which trust is not enough for delegating.

## 2.4. Core beliefs in Trust: non-social Trust

Let's start from *non-social trust*: the most elementary case of trust and delegation is useful in fact to identify what ingredients are really basic in the mental state of trust.

**2.4.1. Trusting objects (non-cognitive and autonomous agents).** What do I have in mind when I trust an object? For example, when I lean against something (ex. a chair, a walking stick), or when I say "I don't trust it" (refusing to lean against it)? what kinds of belief and goal are necessary for trust?

First, I have a **goal** *g* I try to achieve by using *y*: this is what I would like to "delegate to" *y*, its "task" [3]. Then I have some specific beliefs:

1. **"Competence" Belief:** a *positive evaluation* of *y* is necessary, I should believe that *y* is useful for this goal of mine, that it can produce/provide the expected result, that it can play such a role in my plan/action, that it has some function (in the ex. *y* is able to "support" me).

2. **Disposition Belief:** I believe that *y* not only is able and can do that action/task, but it actually will do what I need. With cognitive agents this will be a belief relative to their *willingness*; with objects it is just a belief about their accessibility/non-resistance and their effectiveness. This make them predictable.

3. **Dependence Belief:** Moreover, I should believe -to trust *y* and delegate to it- that either I need it, I depend on it (*strong dependence*; [19], or at least that it is better to me to rely than not to rely on it (*weak dependence*, [20]).<sup>6</sup>

On such a basis I have -when I decide to trust- the new goal that *y* performs  $\alpha$ , and I rely on *y*'s  $\alpha$  in my plan (delegation) (about this decision see later).

Those beliefs define my "trusting *y*" or my **"trust in *y*"**. However another crucial belief arises in my mental state -supported by the previous ones:

4. **Fulfilment Belief:** I believe that *g* will be achieved (thanks to *y* in this case)<sup>7</sup>. This is the **"trust that" *g***. Thus, *when x trusts y for g, it also has some trust that g*.

**2.4.2. Esteem and expectation.** Let us stress the nature of these beliefs about *y*. They are *evaluations* and *positive expectations*, not "neutral" beliefs. In trusting *y* I believe that it has the right qualities, power, ability, competence, and disposition for *p*. Thus, especially among cognitive social agents, the trust that I have in *y* is (clearly and importantly) part of (and is based on) her/his esteem, her/his "image", her/his reputation [21, 22].

We also have a "positive expectation" about *y*'s power and performance. A **positive expectation** is the combination of a belief about the future (prediction) and of a goal: I both believe that *g* and I desire/intend that *g*. In this case: I both believe that *y* can and will do; and I desire/want that *y* can and will do. The fact that when I trust *y* I have 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 me, I cannot do anything else to induce the desired behaviour).

**2.4.3. Trust and Reliance.** On the basis of those beliefs about *y*, *x* "leans against", "depends upon", "relies on", in other words *x* practically "trusts" *y*. Where -notice- "to trust" does not only means those basic beliefs (the core) by 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's call this part "reliance", and the first part "core trust" (and the whole picture mental state of trust and the behaviour "delegation"). We can summarise and simplify the mental ingredients of trust as follows:

GOAL <i>g</i>
B1: <i>y</i> can <i>g</i> , has the power of <i>g</i> (Evaluation)
B2: <i>y</i> will-do for <i>g</i> (Expectation)
B3: <i>g</i> will be true (Trust that <i>g</i> ) <b>CORE TRUST</b>
B4: I need <i>y</i> for <i>g</i> (Dependence)
GOAL of not doing/ not exploit alternatives/ betting on <i>y</i> (Reliance and bet)
GOAL that <i>y</i> can & will do <b>RELIANCE</b>

### Mental ingredients of TRUST

Of course, there is a coherence relation between these two aspects of trust: 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* [23, 24]. The trustful beliefs about *y* (core) are the presupposition of the act of trusting *y*.

<sup>7</sup> 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*.

<sup>4</sup> Suppose for example that you don't trust at all an old wood bridge on a canyon but you are forced to pass over it and thus to rely on it; or 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.

<sup>5</sup> 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.

<sup>6</sup> 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 I trust *y* I feel/believe that it is not harmful, threatening me, dangerous or hostile (for ex. with people I believe that s/he is not an enemy, or has no reason (envy, hate, etc.) to hurt me: diffidence enhances diffidence, sympathy elicits sympathy, trust induces trust in the other.

**2.4.4. Risk, investment and bet.** Any act of trusting and relying implies some bet and some risk [2]. In fact, I might eventually be disappointed, deceived and betrayed by  $y$ : my beliefs may be wrong. At the same time I bet something on  $y$ . First, I renounced to (search for) possible alternatives (for ex. other partners) and I might have lost my opportunity: thus I'm risking on  $y$  the utility of my goal  $g$  (and of my whole plan). Second, I had some cost in evaluating  $y$ , in waiting for its actions, etc. and I wasted my time and my resources. Third, perhaps I had some cost to induce  $y$  to do what I want or to have it at my disposal (for ex. I paid for  $y$  or for its service); now this investment is a real bet [25] on  $y$ .

Thus, to be precise there are two risks:

- a) the risk of failure, the frustration of  $g$  (possibly for ever, and possibly of the entire plan containing  $g$ );
- b) the risk of wasting the efforts.

Not only I risk to miss  $g$  (*missed gains*) but I also risk to waste my investments (*loss*).

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

- *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 [19].

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

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 ex.: I weakly delegate when I bet on an animal in races, or when, being at a bus stop I rely on another person to rise his arm and stop the bus, predicting that he will do this, and risking to miss my 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:** I believe 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.** I should also believe that  $y$  is stable enough in his intentions, that has no serious conflicts about  $\alpha$  (otherwise he might change his mind), or that he is not unpredictable by character, etc.

When I rely on  $y$  for his action, I'm taking advantage of his Independent goals and intentions, predicting his behaviour on such a basis, or I'm myself inducing such goals in order to exploit his behaviour. In any case I not only believe that he is able to do and can do (opportunity), but also that he will do because is committed to this intention or plan (not necessarily to me).

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

Let's *simplify* and formalise this. Using Meyer, van Linder, van der Hoek et al.'s logics [26, 27], and introducing some "ad hoc" predicate (like WillDo, or Persist)<sup>8</sup> we might characterise basic **trust mental state** as follows:

$$\text{Trust}(X,Y,\tau) = K_x \text{Goal}_x(g) \supset B_x \text{PracPoss}_y(\cdot, g) \supset B_x \text{Prefer}_x(\text{Done}_y(\cdot, g), \text{Done}_x(\cdot, g)) \supset (B_x \text{WillDo}_y(\cdot, g) \text{ OR } \text{Goal}_y \text{WillDo}_y(\cdot, g))$$

In other words, the trust is a set of mental attitudes characterizing the "delegating" agent's mind which prefers another agent doing the action.  $Y$  is a non cognitive agent, so  $x$  just believes that  $y$  will do the action. While, if  $y$  is a cognitive agent, the WillDo belief is based on  $y$ 's intention and persistence. Thus, we can shortly characterise **social trust mental state**, i.e. the trust in another intentional agent, as follows:

$$\text{Trust}(X,Y,\tau) = K_x \text{Goal}_x(g) \supset B_x \text{PracPoss}_y(\cdot, g) \supset B_x \text{Prefer}_x(\text{Done}_y(\cdot, g), \text{Done}_x(\cdot, g)) \supset (B_x (\text{Intend}_y(\cdot, g) \supset \text{Persist}_y(\cdot, g)) \text{ or } (\text{Goal}_y (\text{Intend}_y(\cdot, g) \supset \text{Persist}_y(\cdot, g))))$$

$x$  must either believe or have the goal that  $y$  intends to do the action. These are our basic ingredients of social trust.

#### 3.2. 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 [5]  $x$  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, chose, 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; she 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*. I'm so sure that  $y$  will do what I expect (for ex. what he promised) that I do 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 I am deceived and delays possible repairs or protections.

<sup>8</sup> 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.

### 3.3. Trust in strong delegation: the *morality* of *y*

Let's eventually arrive to social trust in strong delegation, which is its typical and strict sense; the sense really considered 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 are hold, but there are some specific additional features.

Not only *y* has an intention to do  $\alpha$ , but he has such an intention (also) because he is "committed to" *x* to do  $\alpha$ ; 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. -[28]) and -in organisations, institution, societies - an obligation (derived from social norms) to do  $\alpha$  (since he promised so to *x*).<sup>9</sup>

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 I support my 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*.

**3.3.1. 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 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 [1] 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*".

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 (as is in fact the book edited by Gambetta), with rationality also "selfishness" and "economic motives" (utility, profit) are smuggled.

When I trust *y* in strong delegation (social commitment by *y*) I'm not assuming that he -by not defeating me- 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.

Thus when I trust *y* I'm just assuming that other motivations will prevail over his economic interests or other selfish goals.

At this level *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*.

This preserve the definition in §3.1 just adding some specification about the motives for *y*'s reliability: the beliefs about *y*'s morality are supports for the beliefs about *y*'s morality or supports for the beliefs about *y*'s intention and persistence. I not only believe that he will intend and persist (and then he will do) but I believe that he will persist *because of certain motives* of his that are more important that other motives inducing him to defection and betrayal. And these motives are already there -in his mind and in our agreement- I have not to find new incentives, to think of additional prizes or of possible punishments. If I'm doing so (for ex. promising or threatening) I don't really trust *y* (yet).

After an agreement we trust *y* because of the advantages we promised (if it is the case), but also or mainly because we believe that he has other important motives like his reputation, or to be honest, or to respect the laws, or to be nice, or to be helpful, etc.

This is the crucial link between "trusting" and the image of "a good person". An honest is an agent who prefers his goal of not cheating and not violating norms to other goals of his such as persuing his own benefits. Social trust is not only a model of *y*'s cognitive and practical capacities, but also of his motives and preferences, and about his morality.

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.

**3.3.2. Trust as a three parties 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 people do not trust each other they 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.

<sup>9</sup> 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.

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 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 (§3.3.1) 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:

$x$  trusts  $A$  and its ability to control, to punish etc. and relies on  $A$  for this;

$x$  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. Notice that  $y$ 's beliefs about  $A$  are precisely  $y$ 's trust in the authority when s/he is the client. When  $y$  is the contractor the same beliefs are the bases of her/his respect/fear toward  $A$  (that is a paradoxical form of trust: trusting a threatening agent!).

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, their 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 which is based on her/his willingness; and reasons for her/his willingness (s/he will do  $\alpha$ , either because of 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. Degrees of Trust: a principled quantification of Trust

The idea that trust is scalable is usual (in common sense, in social sciences, in AI [24, 23]). However, since non 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 strength, 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 the next section 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  $p$* , 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.*

Trust is not for us an arbitrary index just with an operational importance, without a real content.

#### 4.1. Degrees of Trust

Let's call the degree of trust of  $x$  in  $y$  about  $\tau$ :  $\mathbf{DoT}_{XY\tau}$ . 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 - we don't analyse here): the greater  $X$ 's belief in  $y$ ' competence and performance, the greater  $X$ 's trust in  $y$ .

$$\mathbf{DoT}_{XY\tau} = \frac{\text{DoC}_X[\text{Opp}_Y(\cdot, g)] * \text{DoC}_X[\text{Ability}_Y(\cdot)] * \text{DoC}_X[\text{WillDo}_Y(\cdot, g)]}{\text{DoC}_X[\text{WillDo}_Y(\cdot, g)]}$$

where:

- $\text{DoC}_X[\text{Opp}_Y(\cdot, g)]$ , is the degree of credibility of  $X$ 's beliefs about the  $Y$ 's opportunity of performing  $\tau$  to realize  $g$ ;
- $\text{DoC}_X[\text{Ability}_Y(\cdot)]$ , the degree of credibility of  $X$ 's beliefs about the  $Y$ 's ability/competence to perform  $\tau$ ;
- $\text{DoC}_X[\text{WillDo}_Y(\cdot, g)]$ , the degree of credibility of  $X$ 's beliefs about the  $Y$ 's actual performance;

$$\text{DoC}_X[\text{WillDo}_Y(\cdot, g)] = \text{DoC}_X[\text{Intend}_Y(\cdot, g)] * \text{DoC}_X[\text{Persist}_Y(\cdot, g)] \text{ (if } Y \text{ is a cognitive agent)}$$

#### 4.2. To trust or not to trust: degrees of trust and threshold for relying and betting on $y$

Let's now represent a simplified scenario (Fig.1) of the choice of delegating or not [25]:

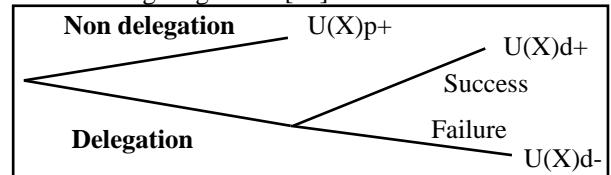


Fig.1

If we call  $U(X)$ , the agent  $X$ 's utility function, we can write:

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

When  $x$  does not delegate he is successfully performing the action by itself:  $U(X)_{p^+}$  is the positive utility of this performance.  $U(X)_{d^+}$  and  $U(X)_{d^-}$  are respectively the

utilities of a successful delegation and the damage of the failure of the delegated agent.

From (1), we can write that for the DoT to be *sufficient for delegating*, the following should hold:

$$\text{DoT}_{XY} > (U(X)_{p^+} - U(X)_{d^-}) / (U(X)_{d^+} - U(X)_{d^-}) \quad (1a)$$

where  $(U(X)_{p^+} - U(X)_{d^-}) / (U(X)_{d^+} - U(X)_{d^-})$  represents the risk factor: *the greater the risk the greater the needed trust*. I.e. the greater the difference between the gain in doing by yourself and the possible loss in delegating, the greater the DoT you need for delegating. Vice versa, the greater the difference between the possible gain and the possible loss in delegating, the smaller the DoT you need for delegating.

Given  $0 < \text{DoT}_{XY} < 1$ , we will have:

$$(U(X)_{p^+} - U(X)_{d^-}) / (U(X)_{d^+} - U(X)_{d^-}) < 1, \text{ from which:}$$

$$U(X)_{d^+} > U(X)_{p^+} \quad (1b)$$

X's DoT in y as for  $\text{DoT}_{XY}$ , can be greater than the risk only if x's possible gain in delegating is greater than x's gain in doing by itself (consider that in this scenario self-confidence is in fact equal to 1: no possible failures!).

A more articulated scenario is outlined in fig. 2.

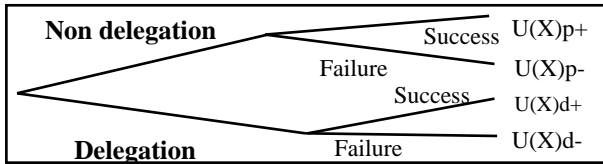


Fig.2

Now there are four possible final situations. We can write:

$$\text{DoT}_{XY} * U(X)_{d^+} + (1 - \text{DoT}_{XY}) U(X)_{d^-} > \text{DoT}_{XX} * U(X)_{p^+} + (1 - \text{DoT}_{XX}) U(X)_{p^-} \quad (2)$$

where  $\text{DoT}_{XX}$  is the *selftrust* of X about  $\text{DoT}_{XX}$  and  $U(X)_{p^-}$  is the utility of the failure of the performance.<sup>10</sup>

Then, we obtain

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

The right part of (2a) represents the new risk factor. Consider that when  $\text{DoT}_{XX} = 1$  (2a) will be reduced to (1a). Let's explain the risk factor by separately considering the two terms of it:

$$A = \text{DoT}_{XX} * (U(X)_{p^+} - U(X)_{p^-}) / (U(X)_{d^+} - U(X)_{d^-})$$

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

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

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  $\text{DoT}_{XX}$ ).

Vice versa, if  $U(X)_{p^+} - U(X)_{p^-} < U(X)_{d^+} - U(X)_{d^-}$

then  $A < \text{DoT}_{XX}$

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  $\text{DoT}_{XX}$ ).

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

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 of non-delegating is bigger than the utility of the failure in delegation, then - in order to delegate - the trust of X in Y about  $\text{DoT}_{XX}$  must be greater than in the case in which the right part of (2a) 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 of non-delegating is smaller than the utility of the failure in delegation, then - in order to delegate - the trust of X in Y about  $\text{DoT}_{XX}$  must be smaller than in the case in which the right part of (2a) is constituted by just A alone.<sup>11</sup>

Since  $\text{DoT}_{XY} < 1$ , from the (2a) we can obtain:

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

From (2b) we can say that, to delegate X to Y the task ( $\text{DoT}_{XY} > \text{risk factor}$ ), as the selftrust ( $\text{DoT}_{XX}$ ) grows, it must also grow the difference between the utility of the success in delegation and the utility of the failure in the non delegation.

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

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

$$U(X)_{d^+} > U(X)_{p^-} \quad (2c)$$

(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.

## 5. 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

<sup>10</sup> More precisely, we have:  $U(X)_{p^+} = \text{Value}(g) + \text{Cost}[\text{Performance}(X)]$ ,  $U(X)_{p^-} = \text{Cost}[\text{Performance}(X)]$ ,

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

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 negative.

<sup>11</sup> 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.

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 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 non-social trust (reliance on objects and tools) and the more complex forms of social trust, based on morality and reputation, have been analysed. We have discussed some unclear 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) [29] of those beliefs.<sup>12</sup>

## 6. References

- [1] D. Gambetta, editor. *Trust*. Basil Blackwell, Oxford, 1990.
- [2] N. Luhmann, Familiarity, confidence, trust: Problems and alternatives. In Diego Gambetta, editor, *Trust*. Chapter 6, pages 94-107. Basil Blackwell, Oxford, 1990.
- [3] Falcone R., A delegation based theory of agents in organizations, *Mathematical Modelling and Scientific Computing*, Vol. 8, 1997 (ISSN 1067-0688).
- [4] 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.
- [5] 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.
- [6] Castelfranchi C., Falcone R., From Task Delegation to Role Delegation, in M Lenzerini (Editor), *AI\*IA97: Advances in Artificial Intelligence*, Lecture Notes in Artificial Intelligence, 1321. Springer-Verlag pg.278-289, 1997.
- [7] P. Maes, Situated agents can have goals. In P. Maes, editor, *Designing Autonomous Agents*, pp. 49-70. The MIT Press, 1990.
- [8] Wooldridge M., and Jennings N., *Intelligent Agents: Theory and Practice*, The knowledge Engineering Review, Vol. 10, N.2, pp. 115-152, 1995.
- [9] Crabtree B., Wiegand M., Davies J., *Building Practical Agent-based Systems*, PAAM Tutorial, London, 1996.
- [10] Cheong, F.C., *Internet Agents: Spiders, Wanderers, Brokers, and Bots*, New Riders Publishing, Indianapolis, USA, 1996.
- [11] Virtual Roundtable, *Internet Computing on-line Journal*, July-August issue, 1997.
- [12] R. Goodwin, Formalizing Properties of Agents. Technical report, CMU-CS-93-159, 1993.
- [13] Winograd, T.A. 1987. Language/Action perspective on the Design of Cooperative Work., In *HCI 3*, 1: 3-30.
- [14] R. Axelrod, *The Evolution of Cooperation*. Penguin Books, London, 1990.
- [15] Castelfranchi, C., Conte, R., Limits of strategic Rationality for Agents and M-A Systems. In *Proceedings of the 4th ModelAge Workshop on "Formal Models of Agents"*, 1997, pp. 59-70.
- [16] Conte, R., Miceli, M., Castelfranchi, C., Limits and Levels of Cooperation. Disentangling various types of prosocial interaction. In Y. Demazeau & J.P. Muller (eds) *Decentralized AI - 2*, Amsterdam, North Holland, 1991.
- [17] Conte, R., Castelfranchi, C., *Cognitive and Social Action*, (Section 10). London, UCL Press, 1995.
- [18] Sichman, J, R. Conte, C. Castelfranchi, Y. Demazeau. A social reasoning mechanism based on dependence networks. In *Proceedings of the 11th ECAI*, 1994.
- [19] P. Dasgupta. Trust as a commodity. In D. Gambetta, editor, *Trust*. Chapter 4, pages 49-72. Basil Blackwell, Oxford, 1990.
- [20] Raub W, Weesie J., Reputation and Efficiency in Social Interactions: an Example of Network Effects. *American Journal of Sociology* **96**: 626-654, 1990.
- [21] S. Marsh, Formalising Trust as a Computational Concept, PhD thesis, Department of Computing Science, University of Stirling, 1994.
- [22] 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.
- [23] M. Deutsch, *The Resolution of Conflict*. Yale University Press, New Haven and London, 1973.
- [24] 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.
- [25] B. van Linder, *Modal Logics for Rational Agents*, PhD thesis, Department of Computing Science, University of Utrecht, 1996.
- [26] Castelfranchi, C., Commitment: from intentions to groups and organizations. In *Proceedings of ICMAS'96*, S.Francisco, June 1996, AAAI-MIT Press.
- [27] R. Demolombe, Formalizing the Reliability of Agent's Information, in *Proceedings of the 4th ModelAge Workshop on "Formal Models of Agents"*, 1997.

---

<sup>12</sup> This work has been realized with contribution of the Joint-Project on "Applicazioni avanzate dell'informatica" (Advanced Application of Computer Science) between Provincia Autonoma di Trento and CNR, 1997 We would like to thank the anonymous ICMAS reviewers for her/his stimulating comments.