Trust and Reciprocity: Misunderstandings

Cristiano Castelfranchi
Trust: Theory and Technology - T[3] Group
Institute for Cognitive Sciences and Technology - CNR - Roma[1]

1. Introduction

This work intends to contest a reductive view of Trust quite diffused in economics, and in psychological and social studies influenced by the Game-Theoretic framework: *the idea that Trust has necessarily to do with contexts which require 'reciprocation'; or that trust is trust in the other's reciprocation*. Trust is even defined in such a way. This is a very reductive view of Trust as mental social attitude, as decision to rely on others, as social relationship.

A multi-layered cognitive model of trust will be proposed. In this model, trust is not conceived only as an *attitude* towards the other, implying different kinds of beliefs (*evaluations*, *expectations*, beliefs on the other's motives, etc.), but also as a *willingness*, a *decision* to rely on the others which makes us dependent and vulnerable from them, as well as a concrete *act* of reliance based on this, and a consequent *social relation* [4] [6] [10].

Also, the concept of Reciprocation/Reciprocity *as behavior and behavioral relation* and the concept of Reciprocation/Reciprocity *as motive and reason for doing something beneficial for the other(s)* will be disentangled. Then, on the base of this conceptual disambiguation it will be argued that we not necessarily trust people because they will be willing to reciprocate; and that we do not necessarily reciprocate for reciprocating.

2. Back to Baruk: Why Trust is not 'benevolence' and 'benevolence' is not trust

"Peace is not the absence of war; it is a virtue, a state of soul. It is a disposition to *benevolence*, to *trust*, to *justice*." (Baruch Spinoza)

In this celebrated sentence Spinoza clearly and directly identifies and distinguishes *the two basic constituents and moves of pro-social relations*:

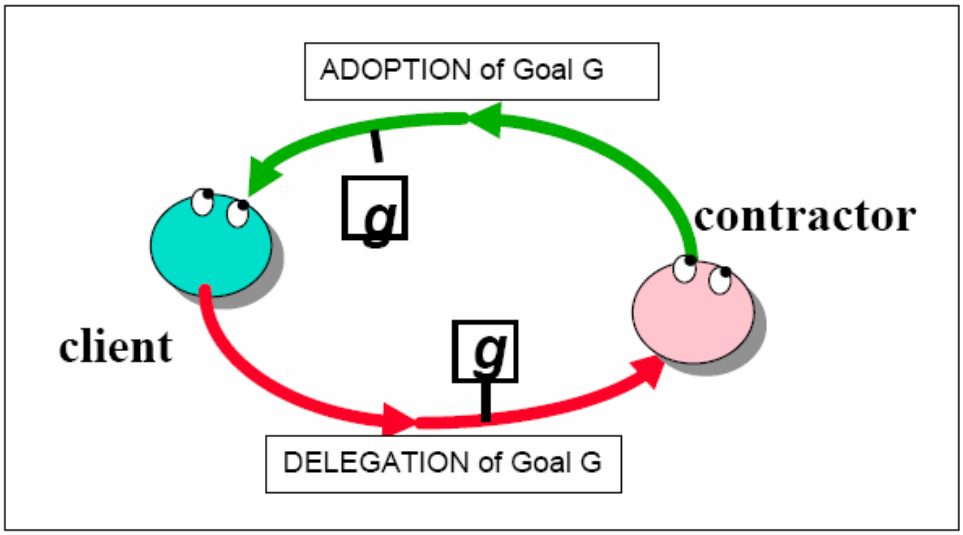
- on one side, goal-adoption, the disposition (and eventually the decision) of doing something *for* the other, of favoring him;
- on the other side, the disposition (and eventually the decision) to count on the other, to delegated to him the realization of our goals, of our welfare [2].

It is important to realize that this basic pro-social structure (the nucleus of cooperation, of exchange, etc.) is *bilateral but not symmetrical.*

Pro-social bilateral relations do not start with 'reciprocation' (which entails some symmetry), with some form of 'exchange'.

The basic structure is composed by a social disposition and act of counting on the other, of being dependent on, of expecting adoption ('*trust*'), to whom hopefully responds a disposition and an act of doing something for the other, of goal-adoption. (Spinoza's '*benevolence*'). [2]

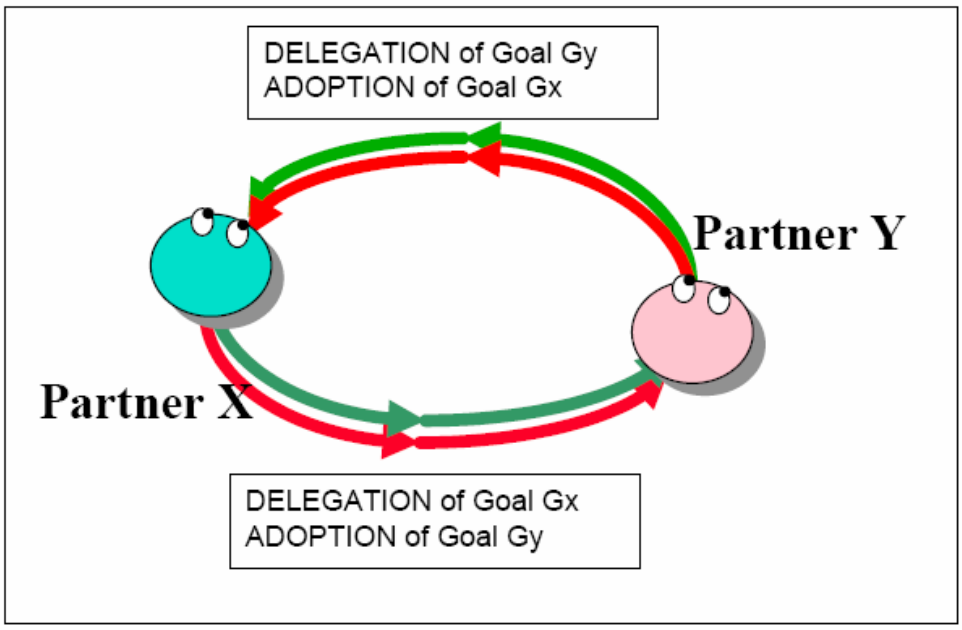
[1] ESF Project "The Social and Mental Dynamics of Cooperation" for the ESF call "The Evolution of Cooperation and Trade" (2007).
I would like to thank Rino Falcone (co-author of our model of trust) and Francesca Marzo for her comments on this paper.



The anti-social corresponding bilateral structure is just: *hostility* (the disposition not to help or to harm) confronting *distrust* and *diffidence*.
'Benevolence' and 'trust' are not at all the same move or disposition (although both are pro-social and can be combined); they belong to and characterize two different although complementary actors and roles.
'Benevolence' and 'trust' – as we have just said - are *complementary* and one related to the other, but they also are in part independent: they can occur alone and can just be 'unilateral'. X can rely on Y, and trust him, without Y being benevolent towards X. Not only in the sense that X's expectation is wrong and she will be disappointed by Y; but in the sense that X can successfully rely on Y and exploit Y's "help" without any awareness or adoption by Y. On the other side, Y can unilaterally adopt X's goals without any expectation from X, and even any awareness of such a help.
Moreover, both trust and 'benevolence' do not necessarily meet and reflect themselves. It is possible an asymmetric trust[3] where only X trusts Y, while Y doesn't trust X (although she knows that X trusts him and X knows that Y doesn't trust her). And this holds both for trust about a specific kind of 'service', and for generalized trust.
Moreover, Trust doesn't presuppose any equality. There can be asymmetric power relationships between the trustor and the trustee: X can have much more power over Y, than Y over X (like in a father-son relation). Analogously, goal-adoption can be fully asymmetrical; where X does something for Y, but not vice versa.

When there is a bilateral, symmetrical, and possibly 'reciprocal' goal-adoption (where the 'help' of X towards Y is (also) due to the help of Y towards X, and vice versa), all that structure of Fig. 1 is doubled. There is trust/reliance from both sides and adoption from both sides.

[2] On the contrary, 'justice' either is the rule for providing adoption or - if interpreted as 'fairness' – it is also the rule of exchange, and presupposes some reciprocity.
[3] This is for example in contrast with May Tuomela's account of Trust [19].



Even in *asynchronous* 'exchanges', even if X acts *before* Y, and Y acts after X's 'help', Y is trusting X. Not necessarily at the very moment of doing his own share, but before, at the very moment of accepting X's help, of relying on it.[4] Of course, in asynchronous 'exchanges' X's trust in Y is broader and more risky: he has additionally to believe (before concrete evidences) that Y will do the expected action, while Y has some evidence of this (but perhaps deceptive).

Trust is not the feeling/disposition of the 'helper' but of the expecting receiver. Trust is the feeling of the helper only if the help (goal-adoption) is *instrumental* to some action by the other (for example, some reciprocation). In this case, X is 'cooperating' towards Y and trusting Y, but because she is expecting something from Y. More precisely (this is the claim that interest the economists) *X is 'cooperating' because she is trusting* (in view of some reciprocation); she wouldn't cooperate without such a trust in Y.[5]
However, as we have explained this is just a very peculiar case; not good at all for founding the notion and the theory of 'trust' and of 'cooperation'.

3. Mixing Up Trust and Goal-Adoption

Let us go deeper in the conceptual problems presented above and let us, for example, consider Yamagishi's interpretation of his comparative results [15] [16] . Following his reasoning, what characterizes Japan is rather "assurance" than true "trust". This means that the Japanese are more "trusting" (α), more dispose to rely on the others, than the Americans, when and if they feel protected by *institutional mechanisms* (authorities and sanctions). Japanese people would tend to "trust" (β) (sic!) only when it is better for them to do so because of the (institutional or social) costs associated with being "untrusting" (sic!); only for avoiding sanctions.
Notice that, first of all, there is a very confuse use of the term "to trust": in the first case (α), it means that X trusts in Y as for doing A, believes that Y is trustworthy and relies on Y; in the second use (β), "to trust" means to contribute, to cooperate! These uses must be distinguished. Obviously they

[4] He has to believe that X's help is good, is as needed, is convenient, is stable enough (it will not be taken back by X), etc.
[5] In other words here we have a *double* and *symmetric* structure (at least in X mind) of goal-adoption and reliance (see later).

are related, since (in Japan) X contributes/cooperates since she worries about institutional sanctions, and she *trusts* the others because she ascribes to the others the same cultural sensibility and worry. But, the two perspectives are very different: expectations about the others' behavior and my own behavior (of contributing) must be distinguished. We cannot simply call "trust" both of them.
Second, the mentioned confusion between "tend to trust" and "tend to cooperate/contribute", and "do not trust" and "do not cooperate/contribute" is misleading per se'. If X cooperates just in order to avoid possible sanctions from the authority or group, trust is not involved. X does not contribute because he trusts or not the others, but for fear of sanctions (X trusts just the authority for monitoring and sanctioning! [4] [10]). Thus calling this cognitive attitude "tendency to trust" is quite confusing.
It is not simply a problem of terminology (and conceptual confusion); it is a problem of *behavioral* notions that are proposed as *psychological* ones. [6]
Finally, here the concept of "(un)trusting" end up loosing his meaning completely. By losing the fundamental ingredients of good evaluating the others, of having good expectations about their behavior, and - for these reasons! – of relying on them and of becoming vulnerable by them), it comes to mean just to 'cooperate' (in game theoretical vocabulary), to contribute to the collective welfare and to risk for whatever reason. The resulting equation "Trust = to contribute/cooperate; untrust = do not contribute/cooperate" is wrong in both the directions; there are behaviors of 'cooperation' without any trust in the others as well as there are trust in the others in non-cooperative situations. First, 'cooperating' for whatever reason is not "to trust". The idea that this *behavior* necessarily denotes 'trust' by the agent and is based on this, and can be used as synonym of it is wrong. For example, as we have already said, worrying institutional sanctions from the authority has nothing to do with trust *in* the other. The problem of confusion between the two attitudes is, between others, due to the fact that, usually, it is not specified "in whom" and "about what", and based on which "expectations and evaluations" about the other a subject trusts. One should be clear and do not confuse between X trusting the others (perhaps because she believes that they worry about the authority and possible sanctions), and X doing something pro-collectivity just because she worries about sanctions, not because she trusts the others. Furthermore, it is wrong that 'trust' coincides with 'cooperating' in those kinds of dilemmas; it is wrong the idea that 'trust' consists in betting on some reciprocation or symmetric behavior. As we will explain, trust holds also in completely different social situations.

4. What trust is and why we decide to trust somebody

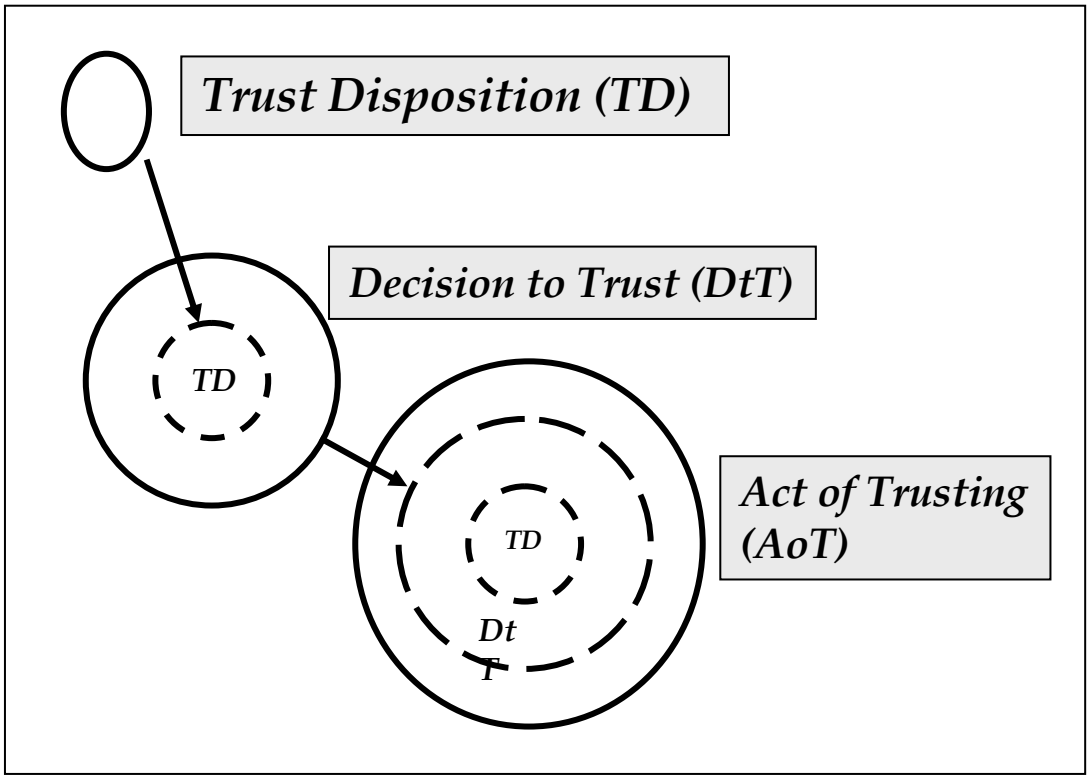
Trust is first of all a disposition, a mental attitude consisting of *beliefs* about the trustee and his behavior.

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

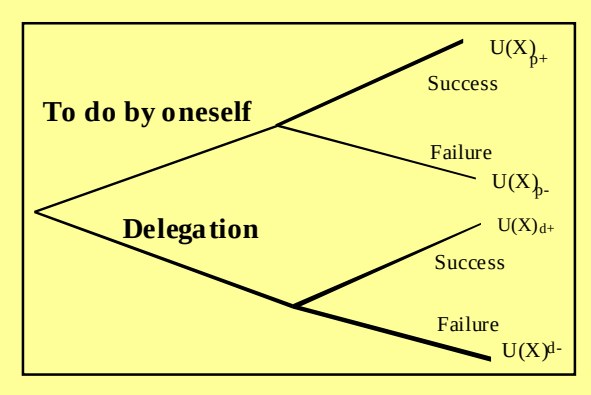
The first (and the third) family of beliefs are '*evaluations*' about Y: to trust Y means to have a good evaluation of him. Trust implies some appraisal.
The second (and the third) family of beliefs are '*expectations*', that is (quite firm) predictions about Y's behavior, relevant for X's goal: X both wishes and forecasts a given action A of Y, and exclude bad actions; she feels safe. [8] [17]

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

The basic nucleus of Trust – as a mental disposition towards Y – is a positive expectation based on a positive evaluation; plus the idea that X would need Y's action.
But trust can be not limited just to a (positive) evaluation, an esteem of Y, and to a potential disposition to relying on him. This potential can become an act. On the basis of such an evaluation and expectations, X can decide to entrust Y within a given 'task', that is to achieve a given goal thank to Y's competent action. 'To trust' is also a *decision* and an *action*. The decision to trust is *the decision to depend on another guy for achieving our own goals*; the intention to rely on the other, to *entrust* the other our welfare.
Thus, we propose a componential and layered model.



Of course what matters is the 'degree' of trust. Is trust both as evaluation and as expectation *enough* for entrusting Y something? For relying and risking on her? How great is the perceived risk (the value of the entrusted/delegated goal plus the possible dangers)? In our model there is a complex decision to trust or not to trust Y as for a given goal/task. This depends not only on the degree of X's trust in Y, but also on the value of the goal; on the perceived risk; on a risk acceptance threshold; etc. [4]



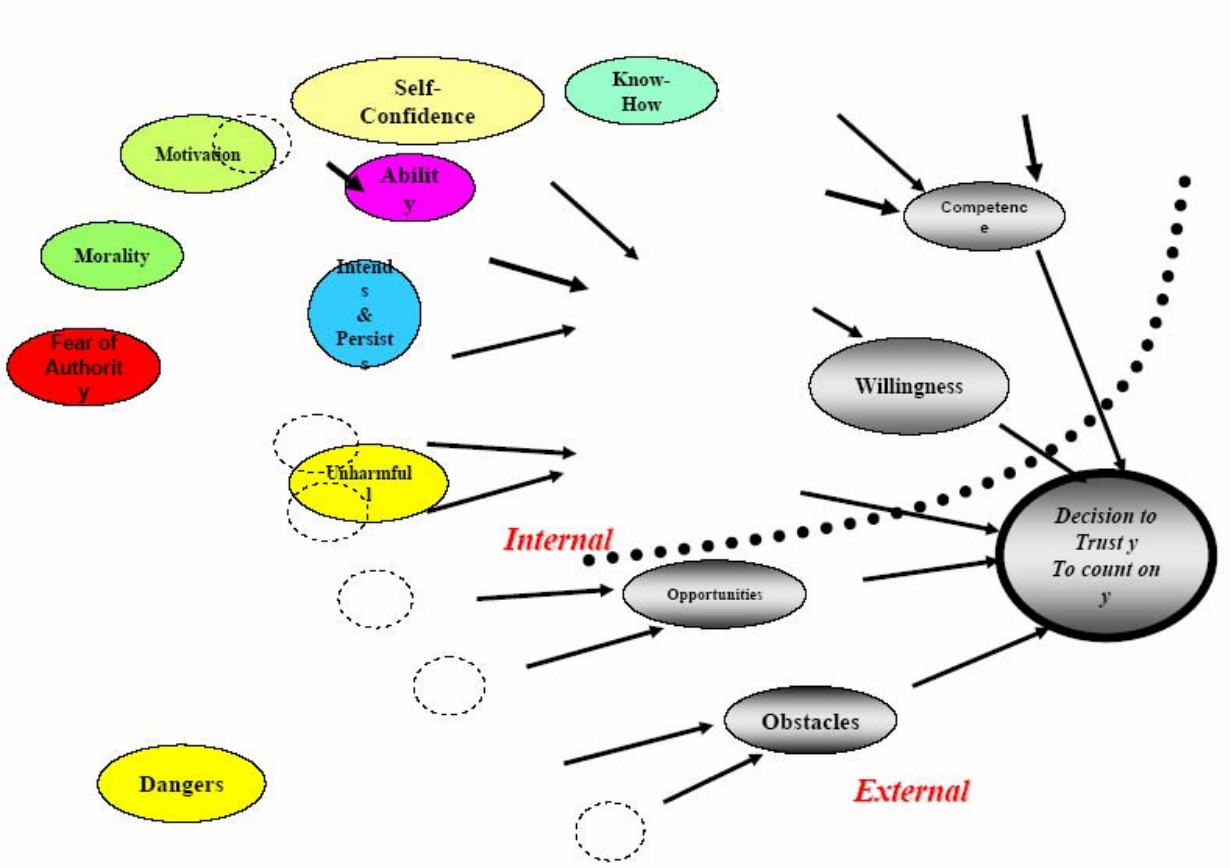
This means that not always and not necessarily we entrust a very trusted guy, or we delegate to the most trusted guy among the possible partners.
But the crucial point is how one can calculate *the degree of trust*. In a belief-based model it derives straight forward from the degree of certainty of the beliefs (and from the possible degree of the 'qualities' of Y).The more I'm sure that Y is (quite) competent, is (quite) able; the more I'm sure that he intends to do the action and will actually do so, the more I trust him for that action.

4.1 Internal vs. external attribution

Of course, it is important also to distinguish between different complementary kinds of trust; since the success of the action doesn't depend only from the agent's competence, skills, and intentional persistence. It also depends on external events (obstacles, opportunities) and infrastructures and tools. Therefore, we have to make another distinction between the Trust in Y (internal attribution), and the trust in favorable circumstances and working infrastructures (external attribution). The decision to trust is due to the global evaluation (internal + external) of trustworthiness; but with various heuristics. Somebody may give more importance to the agent's capacity and willingness (even with adverse circumstances), other will feel safe only in a favorable environment independently on Y's capabilities.
This is also why the idea that Y's failure automatically reduces X's trust in Y, while Y's success necessarily maintains or increase Y's trustworthiness, is wrong. What matters is the 'causal attribution' of the success or failure. If there is an 'internal' (and possibly 'stable') attribution this will impact on Y's trustworthiness (trust in Y), but – if the attribution is 'external' - the success or failure might have no effect at all on Y's trustworthiness. [6] [10]

4.2 Trust is not simply 'subjective probability'

The previous analysis and distinction is one of the reasons why trust cannot be simply reduced to 'subjective probability' of the favorable event/action [12] [20]. Trust in Y is not a mere probability; is an evaluation; and it cannot be mixed up with external circumstances [5]. We have different criteria and different strategies for dealing (both in practice and in our decision making) with these two components.
Moreover, the evaluation beliefs ('competent', 'skilled', 'willing', 'persistent', etc.) on which also the expectation is based, are supported and justified by other belief about Y's features and qualities, which are in fact part of X's trust in Y. Suppose for example that X is sure and feels safe about the fact that Y will do A, just because he has promised to do so, or because this is something fair and morally due to X. This means that X *trusts* Y as a moral guy, as a fair person, as a person keeping promises. These beliefs (this trust relative to 'keeping promises') are the basis and the support for the trust about Y doing A, as promised in this case.



So Trust is a complex picture of Y (mind: including character, motivations, beliefs, feelings, morality, etc.; and body, skills, etc.). Not just a simple and single belief, and even less a simple number.
This is an additional reason why 'subjective probability' is too reductive for representing and accounting for trust [5].

5. A classical definition: What kind of *intention* is 'Trust'?

A part from our socio-cognitive model, let's take a good definition of trust, based on a large interdisciplinary literature and on the identification of fundamental and convergent elements. We will see not only that our model substantially converge with it, but also that, within such a framework, our argument against wrong relationships among trust-cooperation-reciprocity is still valid.

Trust is *"a psychological state of a trustor comprising the intention to accept vulnerability in a situation involving risk, based on positive expectations of the intentions or behavior of the trustee"* (Rousseau et al.) [18]

Notice that "positive expectations of the intentions or behavior of the trustee" do not necessarily refers to an act of reciprocation, but it can be interpreted in a much broader sense. Y can do something good for X – on which X relies – not for reciprocating something done by X. And X on

the other side may trust Y and rely on Y's behavior without having done something for him (like a son towards his parents). Trusting by relying on a reciprocation behavior or motive from Y is just a peculiar sub-case.
This is why it is really misleading to 'define' the act of trusting as an act of 'co-operation', an act of playing a 'cooperative' move in a strategic game. While it would be definitely useful just to claim that this kind of act implies some implicit or explicit trust and that the decision to take a risk is based on a given expectation about Y's intentions or behavior (which entails a positive evaluation of Y).
Both in our model and in this classical definition it is clear that trust is also an 'intention' (and a consequent act); but what kind of intention?
The decision/intention is not about 'doing something <u>for</u> the other', to help or to 'cooperate' with her (in our terminology 'adopting the other's goal'); the act of trusting is not a cooperative act per se. On the contrary, in a certain sense, the trustor X is expecting <u>from</u> the other some sort of 'help' (intentional or non-intentional): an action useful for her.
Of course, in sub-cases, the decision to do something for the other (which is <u>not</u> a decision to trust her) can be joined with and even based on a decision to trust the other, when X is counting on an action of Y useful for herself as a consequence of her own action in favor of Y. One case is in fact when X does something for Y or favoring Y while expecting some reciprocation by Y or for eliciting it.
This is not the only case: X might try to produce an action of Y useful for himself (an action on which she decides to count and bet) not as reciprocation to her 'help', but simply as a behavioral consequence due to Y's independent aims and plans. For example, X might give Y a gun as a gift, because she knows that he hates Z and she wishes that Y will kill Z (not for X but for his own reasons).
Analogously, it is not the case that X always expects an adoptive act <u>from</u> Y and trusts him for this (decides to depend on him for achieving her goal), as 'reciprocation' of her own 'adoption'. However, for sure this is an important family of situations, with various sub-cases quite different from each other from the cognitive point of view.
In some cases, X counts on Y's feeling of gratitude, on a reciprocation motive of affective kind. In other cases she trusts on the contrary Y's interest in future exchanges with her. In others, X relies just on Y's sense of honor and on his sensibility to promises and commitments. In other she knows that Y knows the law and he worries about the authority and its sanctions.[7]
In these cases the act of 'cooperating' (favoring the other and risking on it) is conceived as a (partial) mean for obtaining Y's adoption and/or behavior. Either X wants to provide to Y *conditions* and instruments for his autonomous action based on independent motives, or she wants to provide to Y *motives* for doing the desired action.

5.1 Trusting as a message

In the latter of the cases presented above, it is important to notice that Y has to know (or at least to believe) that X has done a given action *for* him. Thus X (since plans to elicit an adoptive behavior by Y as a specific response to her adoptive act) must ascertain that Y realizes her act toward him and understands its intentional adoptive nature (and the consequent creation of some sort of 'debt'). This means that X's behavior is – towards Y – a 'signal' meaning something to him; in other and better words, it is a form of implicit 'communication' since it is *aimed to* be a signal for him and to mean all that.

[7] Notice that X might also adopt Y's goals, while expecting for her 'cooperation', not as a means for this. X might for example be an anticipatory reciprocator; since he knows that Y is doing an act in favor of him, he wants to reciprocate and – in advance – does something for Y

X's cooperation in view of some form of an intentional reciprocation (of any kind) need to be a *behavioral implicit communication act* because Y's understanding of the act is crucial for providing the right motive for reciprocating.
This doesn't mean that necessarily X intends that Y understands that she intends to communicate (gricean meta-message): this case is possible and usual but not necessary. Let us suppose, for example, that X desires some favor from Y and in order to elicit a reciprocating attitude does something in favor of Y (say, a gift). It is not necessary (and sometimes is even counterproductive) that Y realizes the selfish plan of X, and thus the fact that she wants that he realizes that she is doing something 'for' him and *intends that he recognizes this.* It is sufficient and necessary that Y realizes that X is intentionally doing something just for him, and for sure X's act is also aimed at such recognition by Y: X's intention to favor Y must be recognized, but X's intention that Y's recognizes this doesn't need to be recognized. [3]

6. Does Trust presuppose Reciprocity?

Following the model introduced in the previous sections, it is possible to contradict a typical unprincipled and arbitrary restriction of the notion and of the theory of trust, present in some of economic-like approaches. It is based on a restriction of trust only to exchange relations, in contexts implying *reciprocity*. It is, of course, perfectly legitimated and acceptable to be interested in a sub-domain of the broad domain of trust (say "trust in exchange relations"), and to propose and use a (sub-)notion of trust limited to those contexts and cases (possibly coherent or at least compatible with a more general notion of trust). What would be less acceptable is proposing a restricted notion of something - fitting with peculiar frame and specific issues - as the only one, generally valid.
Consider, by way of example, one of these limited kind of definition, clearly game theory inspired, and proposed by R. Kurzban [14]: trust is *"the willingness to enter exchanges in which one incurs a cost without the other already having done so"*.
As we have seen the most important and basic constituents of the mental attitude underlying trust behavior are already present (and more clear) in *non-exchange* situations.
Y can do an action in favor of X for a lot of motives not including reciprocation; analogously, X can rely on Y's action for a broad set of different motives ascribed to Y (for instance, friendship, honesty, generosity, search for admiration, etc.) and the reasons active in cooperation, exchange, reciprocation situations, are only a subset of them.
It is simply false that we feel trust or not, and we have to decide to trust or not, only in contexts of reciprocation, when we do something for the other or give something to the other and expect (wish) that the other will reciprocate doing his share, and will not defeat. This notion of Trust is arbitrarily restricted and cannot be useful to give an account to, the case where Y simply and unilaterally offers and promises to X that she will do a given action A for him, and X decides to count on Y, does not commit herself to perform A, and *trusts* Y for accomplishing the task. The very notion of trust must include cases like this that describe real life situations quite relevant in society. Should we even search just for a 'behavioral' notion, *doing nothing* and counting on others is a behavior.

At this point, it is important to remind that trust can be just unilateral or even based on Y's ignorance: X can trust Y as for doing a needed action and relying (risking/betting) on this, but, perhaps, Y ignores that X is exploiting him.
Even cases based on an explicit agreement do not necessarily require reciprocation. Consider a real life situation where X asks to Y "Could you please say this to the Director, when you see her; I have no time; I'm leaving now", she is in fact trusting Y as for really doing the required action. Y is expected to do this not for any reciprocation (but, say, for courtesy, friendship, pity, altruism, etc.).

One might claim that X has given something to Y: his gentle "Please"; and Y has to do the required action in order to reciprocate the "Please". But this is frequently not true since this is usually not enough for doing: it is not the reason why X expects that Y will do so (in fact X has to be grateful after the action and she is in debt); it is not what Y feels and the reason why he does the action; he feels that his cost exceeds a lot the received homage. Moreover, there might be other kinds of requests, based on authority, hierarchy, etc. when X doesn't give to Y anything at all 'in change' of the required action which is simply 'due'. But also in these cases X considers Y trustworthy if is relying on him.
In sum: *trust is not an expectation of reciprocation; and doesn't apply only to reciprocation situations*.
Related to this misunderstanding is the fact that "being vulnerable" is often considered as strictly connected with "anticipating costs".
This diffuse view is quite coarse: it mixes up a correct idea (the fact that trust - as decision and action - *implies a bet, taking some risk, be vulnerable* [1] [13] [18] with the reductive idea of *an anticipated cost, a unilateral contribution*. But in fact, to contribute, to 'pay' something in anticipation while betting on some 'reciprocation', is just one case of taking risks. The expected beneficial action ("on which our welfare depends" [1] [13] [18] from the other is not necessary 'in exchange'.
The risk we are exposed to and we accept when we decide to trust somebody, to rely and depend on her, is not always the risk of wasting our invested resources, our 'anticipated costs'. The main risk is the risk of not achieving our goal, of being disappointed as for the entrusted/delegated and needed action, although perhaps our costs are very limited (just a verbal request) or nothing (just exploiting her independent action and coordinate our own behavior). Sometimes, there is the risk of frustrating forever our goal since our choice of Y makes inaccessible other alternatives that were present at the moment of our decision. We also risk the possible frustration of other goals: for example, our self-esteem as good and prudent evaluator; or our social image; or other goods of us that we didn't protect from Y's access. Thus, it is very reductive to identify the risks of trust with the lack of reciprocation and thus wasting our investment; risk, which is neither sufficient nor necessary.

6.1 Reciprocation vs. Reciprocity

In order to understand the relationship between trust and reciprocity, together with a correct theory of trust, it is necessary to have a better theory of what it means to reciprocate as well. Although we do not intend to present it here, we want to propose a basic distinction necessary for this kind of theory: it is important to distinguish between Reciprocation - as a behavior - and Reciprocity as an attitude and a motivation [9].
The two of them not necessary co-occur: not all cases of Reciprocation are due to the motivation of Reciprocating. X might 'contribute to the other welfare' (doing something for Y or for the collective) - when in fact the other already did so and thus X's action 'reciprocates' Y's action - without even realizing of being reciprocating, but for his own motives (pro-social or others). Or X might be aware of being actually 'reciprocating' but doing this not for Reciprocity, but for social or legal norms (for example, for conformity or for avoiding bad reputation or other sanctions).

- *Reciprocity is only one among many possible motives on which we rely when trusting another agent as for doing the desired and expected behavior.*
- *Reciprocity is only one among many possible motives inducing to do our share (expected by the other) after the other has trusted us.*

In asymmetric Trust action, for example, where only Y has to do something 'for X' (even possibly ignoring that X is relying on him and exploiting him), X is for sure not counting on Y's reciprocity motive.

Even in situation of exchange or cooperation, where X has done something for Y (of for the common goal) and is expecting that Y does the same, not necessarily she is counting on Y's reciprocation due to a Reciprocity motivation.
One might claim that the trusting attitude of X towards Y *per se* elicits and counts on Y's reciprocation of trust and a reciprocity tendency (see below). This happens frequently but is not necessarily true.
Even when Y knows that we trust her and have delegated to her making us vulnerable from her, not necessarily she will do the expected action *for reciprocating our trust* (which is not necessarily particularly 'beneficial' and useful for her). She can do the action for many reasons: pity, altruism, obligation (we trust for example policemen for doing what is in their role), avoiding threatened harms, getting a reward, etc. Reciprocity of trust is neither a sufficient motive (otherwise any reward or sanction would be useless) nor a necessary motive. In addition, trust does not usually rely on Reciprocity motives (even in reciprocation relations like exchange) also because Reciprocation too can be not grounded on and motivated by Reciprocity.

7. Relations between Trust and Reciprocity

Given that Trust doesn't necessarily presuppose reciprocity, and vice versa, let's now give a look to several interesting and principled relationships between Trust and Reciprocity. Let's consider some of them:

(i) X trusts that Y will reciprocate his adoption.
He is *betting* on Y reciprocation.

(ii) If Y reciprocates, X will trust her next time.

X uses Y's behavior as a *sign* or *signal* of some reasonably stable disposition (towards her), as an evidence for future behaviors.

The more Y's act is costly for Y, and the more Y would have been in condition of safely not reciprocating, the more the signal is reliable. The credibility of the signal is function of Y's cost * impunity (the probability of not being detected or punished).
Y's behavior is also a very good sign of his adoptive attitude and thus of his trustworthiness (and thus produces not only gratitude but trust for the future) when Y is 'generous' in giving, that is, he is giving even if it is not 'due', or more than 'due'. The more generous Y, the more credible his behavior as *sign* of his 'benevolence'.

(iii) Y reciprocates in order X trusts him next time (and exchange with him, chose him as partner).
That is Y is paying a cost for acquiring some trust-capital [7]; it is an investment for future exchanges.

(iv) Trust can be about trust, and about a reciprocation of trust
A very remarkable case is when not only trust faces trust (T < == > T) but one is about the other; when X's trust is about Y trusting him.
There is some sort of "meta-trust": "I trust Y, since she trusts me" (and vice versa). Many social exchanges require this form of mutual and meta Trust. And in those cases X trusts Y also *in order* Y trusts X; more precisely "since and in order to"; it is some sort of self-fulfilling prophecy.

7.1 Does Trust create Trust and there exists a Norm of reciprocating Trust?

We have made clear that it is not for reciprocation that Y does the expected action after we trust him and decided to rely and depend on him; and also that trust is not always 'reciprocated' (even when Y performs the entrusted action). However, we acknowledge that there exist a property of trust to elicit trust, and we wonder about the idea that it might even exist *a norm of trust reciprocation*.
Since trust is not just a behavior, but a mental state and a feeling, it cannot be really 'prescribed', since is not really 'voluntary'. Only the act, the intention can paradoxically be 'prescribed': "Trust him! Rely on him!"; but not the real background disposition. [8]
However, moral (ad religious) norm can impinge even on mere mental dispositions ("Do not desire…" "Do not have this kind of thoughts!"); thus there might be, and in fact it seems that there is, a social-moral *norm* about reciprocating Trust: "Since/if X trusted you, you have to trust X". To trust somebody seems to be a form of 'gentle' disposition or act, and it seems that we have to respond to a gentle act with a gentle act, to a smile with a smile.
There is a clear psychosocial phenomenon of trust propagation such that trust creates trust while diffidence creates hostility. If X trust Y, this tends to elicit not only a 'benevolent' but also a 'trustful' attitude in Y towards X. However, we do not believe that it is mainly due to such a possible moral norm. We believe that it is mainly due to:

- the fact that while trusting Y, X makes himself dependent and vulnerable from Y, more exposed , and thus less dangerous, harmless;
- the fact that while trusting Y, X shows to have positive evaluations, esteem, thus a good disposition towards Y, which can be a good basis and a prognostic sign for 'benevolence' towards Y, that is, for adoption; (it is more probable that we help somebody that we perceive as competent and benevolent, although we do not currently intend to exchange with him);
- the fact that while trusting Y, X may even rely on common values, on sympathy (common feelings), on a sense of common membership, etc. and this makes him at his turn reliable, safe.

Nevertheless, we believe that such a Norm of responding with trust to trust, exists. It is not importantly responsible for eliciting trust in response to trust, but is important for other functions. It is used for moral *evaluation*, and is responsible of blame, shame, etc..

8. Concluding remarks

We have argued against *the idea that Trust has necessarily to do with contexts which require 'reciprocation'; or that trust is trust in the other's reciprocation*.
A multi-layered cognitive model of trust has been presented, to argue against too reductive views (like the identification of trust with the subjective probability of the favorable event). In this model, trust is not conceived only as an *attitude* towards the other, implying different kinds of beliefs (*evaluations*, *expectations*, beliefs on the other's motives, etc.), but also as a *willingness*, a *decision* to rely on the others which makes us dependent and vulnerable from them, as well as a concrete *act* of reliance based on this, and a consequent *social relation*.
We have also introduced a distinction between, the concept of Reciprocation/Reciprocity *as behavior and behavioral relation* and the concept of Reciprocation/Reciprocity *as motive and reason for doing something beneficial for the other(s)* will be disentangled.
On the base of this conceptual disambiguation it has been argued that we not necessarily trust people because they will be willing to reciprocate; and that we do not necessarily reciprocate for reciprocating.

[8] In those extreme cases Trust as disposition wouldn't be enough for the intention, but we add independent, external, additional reasons which forces us to 'trust' in the sense of deciding to rely on Y.

References

[1] Barber, B. (1983). The Logic and Limits of Trust. New Brunswick, NJ: Rutgers University Press.

[2] Castelfranchi, C. (1998). Modelling Social Action for AI Agents. *Artificial Intelligence*, 103, 1998, 157-182.

[3] Castelfranchi C. (2004). Silent Agents. From Observation to Tacit Communication. Modelling Other Agents from Observations: MOO 2004 A one-day workshop to be held as part of the International Joint Conference on Autonomous Agents and Multi-Agent Systems July 19, 2004 URL: http://www.cs.biu.ac.il/~galk/moo2004/

[4] 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, pp.72-79.

[5] Castelfranchi C., Falcone R. Trust is much more than subjective probability: Mental components and sources of trust, *32nd Hawaii International Conference on System Sciences* - Track on *Software Agents*, Maui, Hawaii, 5-8 January 2000. Electronic Proceedings

[6] Castelfranchi, C. and Falcone, R. "Trust Theory" . John Wiley & Sons, UK (in preparation)

[7] Castelfranchi, C. , Falcone, R., Marzo, F. (2006), Being Trusted in a Social Network: Trust as Relational Capital. Proceedings of *iTrust 2006* - 4th International Conference on Trust Management, Pisa, 16-19 May, 2006.

[8] Castelfranchi, C., Lorini, E. (2003). Cognitive Anatomy and Functions of Expectations. In *Proceedings of IJCAI'03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*, Acapulco, Mexico, August 9-11, 2003.

[9] Cialdini, R. B. (2001). *Influence: Science and practice* (4th ed.). Boston: Allyn & Bacon.

[10] Falcone R., Castelfranchi C., (2001). The Socio-cognitive Dynamis of Trust: Does Trust Create Trust? In R. Falcone, M. Singh, Y.H. Tan (eds.) *Trust in Cyber-societies. Integrating the Human and Artificial Perspectives*. Heidelbergh, Springer, LNAI 2246, pp. 55-72.Kurzban, R. (2001) Are experimental economics behaviorists and is behaviorism for the birds? Behavioral and Brain Sciences, 24, pp. 420-1

[11] Fukuyama, F. (1995). Trust: The Social Virtues and the Creation of Prosperity. New York: The Free Press

[12] Gambetta, D. (ed.) (1988). Trust: Making and Breaking Cooperative Relations. New York: Basil Blackwell.

[13] Hardin, R. (2004). Trust and Trustworthiness. New York: Russel Sage Foundation.

[14] Kurzban, R. (2003) Biological foundation of reciprocity. In E. Omstrom and J. Walker (eds.) *Trust, Reciprocity: Interdisciplinary Lessons from Experimental Research* (pp. 105-27). N.Y., Sage.

[15] Yamagishi, T. (2003) Corss-societal experimentation on trust: A comparison of the United States and Japan. In E. Omstrom and J. Walker (eds.) *Trust, Reciprocity: Interdisciplinary Lessons from Experimental Research* (pp. 352-70). N.Y., Sage.

[16] Mashima,R. Yamagishi, T. and Macy, M. (2004) Trust and cooperation: A comparison between Americans and Japanese about in-group preference and trust behavior. Japanese Journal of Psychology, 75, pp. 308-15.

[17] Miceli, M. e Castelfranchi, C. (2002). The mind and the future: The (negative) power of expectations. *Theory & Psychology*. Vol.12 (3): 335-.366

[18] Rousseau, D.M., Burt, R.S., Camerer, C. (1998) Not so different after all: a cross-discipline view of trust. *Journal of academy management review*. 23 (3), 393-404.

[19] Tuomela, M. Trust and Collectives. Doctoral dissertation, September 2006. University of Helsinki, Faculty of Social Sciences, Department of Social Psychology; http://ethesis.helsinki.fi/julkaisut/val/sosps/vk/tuomela/

[20] Williamson, O. (1985)*The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. New York: The Free Press, 1985.