Int. Rev. Econ. (2008) 55:45–63
DOI 10.1007/s12232-008-0041-x

Trust and reciprocity: misunderstandings

Cristiano Castelfranchi

Received: 15 October 2007 / Accepted: 30 January 2008 / Published online: 21 February 2008

Abstract We contest a reductive view of trust, quite diffused in economics, and in studies influenced by the Game-Theory framework: the idea that trust has necessarily to do with contexts requiring "reciprocation"; or that trust is trust in the other's reciprocation. A multi-layer cognitive model of trust will be proposed. 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 to rely on others that makes us dependent and vulnerable to them, as well as a concrete act of reliance based on this. Not necessarily we trust people because they will be willing to reciprocate; and we do not necessarily reciprocate for reciprocating. Trust (even "genuine" trust) is based on a variety of motivations ascribed to others and makes prevail the adoption of our needs and goals: from "altruism" to "self-interest", from reciprocation to norms or to affective reasons.

Keywords Trust · Reciprocity · Cognitive theory · Goal-adoption

JEL Classifications A12 · A13 · C79 · D8

C. Castelfranchi (✉)
Institute of Cognitive Sciences and Technologies, National Research Council,
Via San Martino della Battaglia 44, 00185 Rome, Italy
e-mail: cristiano.castelfranchi@istc.cnr.it

C. Castelfranchi
Department of Communication Sciences, University of Siena, Siena, Italy

1 Why "trust" is not "benevolence" or "cooperation", and "benevolence" is not trust

In social theory we have to clearly distinguish *the two basic constituents and moves of pro-social relations*[1]:

- 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 delegate him for the realization of our goals, of our welfare (Castelfranchi 1998).

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 by the act of counting on the other, of being dependent on, of expecting adoption ("trust"); while, on the other side, it hopefully responds to that disposition and to an act of doing something for the other, an act of goal-adoption (Spinoza's "benevolence").[2] "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 partially 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, trust and "benevolence" do not necessarily meet and reflect themselves. An asymmetric trust is possible, where only X trusts Y, while Y doesn't trust X (although he knows that X trusts him and X knows that Y doesn't trust her). Analogously, trust doesn't presuppose any equality. There can be asymmetric power relationships between the trustor and the trustee: Y can have much more power over X, than X over Y (like in a father–son relation), or vice versa.

When there is a bilateral, symmetrical, and possibly "reciprocal" goal-adoption (where the "help" of X to Y is (also) due to the help of Y to X, and vice versa) there is *trust/reliance* from both sides and *adoption* from both sides.[3]

[1] "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).

[2] The anti-social corresponding bilateral structure is just: *hostility* (the disposition not to help or even to do harm) confronting *distrust* and *diffidence*.

[3] 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, and relying on it. Of course, in asynchronous "exchanges" X's trust in Y is broader and more risky: she has additionally to believe (before concrete evidences) that Y will do the expected action, while Y has some evidence of this (but perhaps deceptive).

In sum, 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" with Y and trusting Y, but only because she is expecting something from Y. More precisely (this is the interesting claim for the economists) *X is "cooperating" because she is trusting* (in view of some reciprocation); she wouldn't cooperate without such a trust in Y.[4]

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".

Let us also consider—before introducing our complex model of trust—a good definition of trust, based on a large interdisciplinary literature and on the identification of fundamental and convergent elements.

Trust is "a psychological state of a trustor comprising the intention to accept vulnerability in a situation involving risk, based on positive expectations of the intentions or behavior of the trustee" (Rousseau et al. 1998).

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

Notice that the *decision/intention* is not about "doing something *for* the other", to help or to "cooperate" with him (in our terminology "adopting the other's goal"); the act of trusting is not a cooperative act per sè. On the contrary, in a certain sense, the trustor X is trying to exploit the other, is expecting *from* the other some sort of "help" (intentional or non-intentional).

Of course, in sub-cases, the decision to do something for the other (which is *not* a decision to trust him) 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 while expecting or eliciting some reciprocation from Y.

This is not the only case of active *influencing* (manipulation) Y's behavior in order to obtain the desired action: X might try to produce an action of Y that is useful for herself (an action on which she decides to count and bet) not as a 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 Y to kill Z (not for X but for his own reasons).

Analogously, it is not the case that X always expects an adoptive act *from* Y and trusts him for this (decides to depend on him for achieving her goal), as a "reciprocation" of her own "adoption". However, for sure this is an important

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

family of situations, with various sub-cases, one 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 other cases, X relies just on Y's sense of honor and on his sensibility to promises and commitments. In other cases again, she knows that Y knows the law and worries about the authority and its sanctions.[5]

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 Y with *conditions* and instruments for his autonomous action based on independent motives, or she wants to provide Y with *motives* for doing the desired action.

2 Mixing up trust and goal-adoption

Let us go in depth in the conceptual problems presented above and let us, for example, consider Yamagishi's interpretation of his comparative results (Yamagishi 2003; Mashima et al. 2004). Following his reasoning, what characterizes Japan is rather "assurance" than true "trust". This means that the Japanese are more "trusting" (a), more disposed to rely on others, than the Americans, when and if they feel protected by *institutional mechanisms* (authorities and sanctions). Japanese people would tend to "trust", (b) (cooperate) (sic!) only when it is better for them to do so because of the (institutional or social) costs associated with being "untrusting" (sic!); only to avoid sanctions.

Notice that, first of all, there is a very confused use of the term "to trust": in the first case (a), it means that X trusts that Y will do *A*; X believes that Y is trustworthy and relies on his realizing some goal; X is waiting for Y's action. In the second case (b), "to trust" means to contribute, to cooperate, to *do* something for the others! These cases must be distinguished. Obviously they are related, since (in Japan) X contributes/cooperates because she worries about institutional sanctions, and she *trusts* the others because she ascribes them 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 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 sè. If X cooperates just in order to avoid possible sanctions from the authority or group, trust is not necessarily involved. X does not contribute because she trusts or not the others, but for fear of sanctions. X trusts just the authority for monitoring and sanctioning! (Castelfranchi and Falcone 1998; Falcone and Castelfranchi 2001). Thus calling this cognitive attitude "tendency to *trust*" is quite confusing.

[5] Notice that X might also adopt Y's goals, while expecting his "cooperation", not as a means for this. X might for example be an anticipatory reciprocator since she knows that Y is doing an act in her favor, she wants to reciprocate and—in advance—she does something for Y.

Finally, here the concept of "(un)trusting" ends up losing its meaning. It loses the fundamental ingredients of a positive evaluation of the other, of a positive expectation about his behavior, and—for these reasons!—of the decision to rely on him and to become vulnerable to him, and 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 = not to contribute/cooperate" is wrong in both directions; there are behaviors of "cooperation" without any trust in the others as well as there is 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 its synonym, is wrong. For example, as we already said, worrying about institutional sanctions from the authority has nothing to do with trust *in* the others. The problem of confusion between the two attitudes is, among others, due to the fact that, usually, it is not specified "in whom" and "about what", and is not based on which "expectations and evaluations" about the others.

Let us give another significant example of this deformed view of trust. Consider the definition, clearly inspired by Game Theory, proposed by Kurzban (2003): trust is "the willingness to enter exchanges in which one incurs a cost without the other already having done so".

It is simply false that we feel trust or not, and we have to decide to trust or not only in contexts of exchange and 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 account for the case where Y simply and unilaterally offers and promises X to do a given action *A* for her, and X decides to count on Y, does not commit herself to perform *A*, and *trusts* Y to accomplish 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 passive "behavioral" notion? *Doing nothing* and counting on others is a behavior.

In sum: *trust is not an expectation of reciprocation; and doesn't apply only to reciprocation situations*.

The fact that "being vulnerable" is often considered as strictly connected with "anticipating costs" is related to this misunderstanding. This widespread 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* (Luhman 1979, 1990; Barber 1983; Rousseau et al. 1998; Gambetta 1988), with the reductive idea of *an anticipated cost, a unilateral contribution*. But in fact, to contribute, to "pay" something in advance while betting on some "reciprocation", is just one case of taking risks. The expected beneficial action from the other ("on which our welfare depends") is not necessary "in exchange" (Hardin 2002). The risk we are exposed to and we accept when we decide to trust somebody, to rely and depend on him, 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 for the entrusted/delegated and needed action, although perhaps our costs are very limited (just a verbal request) or nothing (just exploiting his 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 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 the waste of our investment.

3 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.

(1) X believes that Y is able and well disposed (willing) to do the needed action.
(2) X believes that in fact Y will appropriately do the action, as she hopes.
(3) 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 excludes bad actions; she feels safe (Miceli and Castelfranchi 2002; Castelfranchi and Lorini 2003).

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 needs or might 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 rely on him. This potential can become an act. On the basis of such an evaluation and expectation, X can decide to entrust Y with a given "task", that is to achieve a given goal thanks 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 to achieve our own goals*; the intention to rely on the other, to *entrust* the other with our welfare (Fig. 1).

Thus, we propose a componential and layered model (Table 1). There is a central kernel or core of trust attitude including X's goal *p*, and several beliefs of X: the belief that Y would be able and in condition to realize such a goal (positive

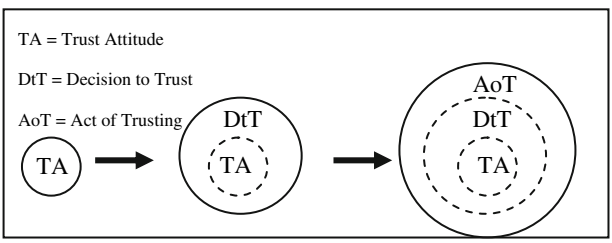

Fig. 1 Trust layers

evaluation of Y); the belief that Y would actually do the needed action *A* for realizing *p* (positive expectation); the belief that X would need Y's action, depends on Y; the belief that *p* would be realized. Not only X trusts Y, or trusts "in" Y (as for performing the action); but—consequently—X also trusts that *p* will be realized. When X—on such a basis—also arrives to the "decision" to rely on Y, there are additional mental ingredients; at least: the decision and thus the goal of betting on Y and not personally pursuing the goal, and the goal (not only the "prediction") that Y is able and will actually do the needed action.

What certainly matters is the "degree" of trust. Is trust, both as evaluation and as expectation, *enough* for entrusting Y with something? To rely and risk on him? 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 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 (Fig. 2); on a risk acceptance threshold; etc. (Castelfranchi and Falcone 1998).

This means that not always and not necessarily we entrust a very trustworthy guy, or we delegate the most trustworthy guy among the possible partners; it depends on costs, etc.

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.

3.1 Internal versus external attribution

Certainly, it is also important to distinguish among different complementary kinds of trust, since the success of the action does not depend only on the agent's competence, skills, and intentional persistence. It also depends on external events

Table 1 Mental ingredients of trust

Goal-1: p	
Belief-1: Y can realize p; has the power of realizing p	Core trust
(Evaluation)	
Belief-2: I need Y for achieving p	
(Dependence)	
Belief-3: Y will do action A for p	
(Expectation)	
Belief-4: p will be realized	
(Trust "that p will be realized")	
Goal-2: not doing A/not exploiting alternatives/betting on Y	Reliance
(Reliance and bet)	
Goal-3: Y can & will do A and realize p	

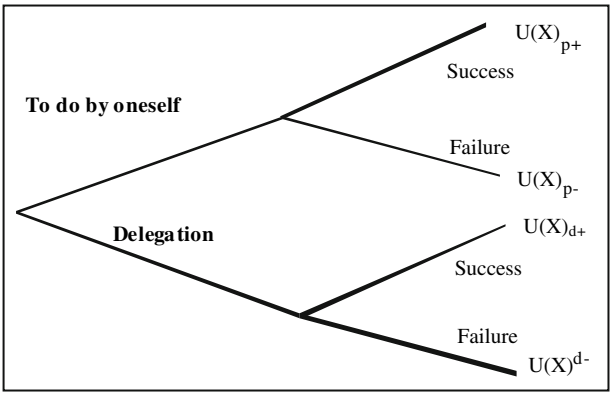

Fig. 2 The decision scenario

(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 ability and willingness (even with adverse circumstances), while others will feel safe only in a favorable environment independently of Y's skills.

This is also the reason why the idea that Y's failure automatically reduces X's trust in Y, while Y's success necessarily maintains or increases Y's perceived 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" (just due to the circumstances)—the success or failure might have no effect at all on Y's perceived trustworthiness (Falcone and Castelfranchi 2001; Castelfranchi and Falcone 2008).

3.2 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 (Gambetta 1988; Willianson 1985, 1993); a frequent reductive view in economists. Trust ***in*** Y is not a mere probability; it is an evaluation; and it cannot be mixed up with external circumstances. 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 beliefs 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. But they also are—recursively—sub-kinds of trust: "I trust (in) Y's morality", "I trust (in) Y's fear of bad reputation", "I trust (in) Y's expertise". 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 (Castelfranchi and Falcone 2000).

Actually, we do agree with Williamson (1993): if trust is just a euphemism for "subjective probability" we do not need it, we already have a strong theory of it, and a new vague and merely suggestive term is just confusing cosmetics. On the contrary, we believe that trust is a specific, well-defined, mental and social construct, and in our social decisions we build on trust not simply on regularities and probability.

3.3 "Genuine" social trust: trust in "adoption"

To be true, trust is not only a "social" attitude. It can be directed towards an artifact or unanimated process.[6] However, it is true that the most theoretically and practically relevant and the most typical notion of trust is the social one.

"Social" trust means trust in another *autonomous* agent, considered as such, with its attitudes, motivations (including the social ones), and some freedom of choice. It requires an "intentional stance" towards a social entity (with its own intentional stance towards us).

This is not yet enough for capturing the most *typical* social notion of trust; what many authors (like Baier 1986; Hardin 2002; Holton 1994, and others) would like to call "genuine" trust.

"Genuine" (social) trust, the basic, natural form of social trust, is based on Y's "adoptive" attitude. That is, X trusts Y's adoption of her interest/goal, and counts on this.

Y is perceived as taking into account X's goals/interests; and possibly as giving priority to them (in case of conflicts). This is true trust in a social agent "as a social agent".

"Social goal-adoption" is the idea that another agent takes into account in his "mind"—in order to satisfy them—my goals (needs, desires, interests, projects, etc.); he "adopts" them as his own goals, since he is an "autonomous agent", i.e., self-driven and self-motivated (but not necessarily "selfish"!), and he is not a hetero-directed agent, and can only act in view of, be driven by, some *internal* purposive representation. So—if such an (internally represented) goal will be preferred to others—he will be regulated by my goal; for some motive he will act in order to realize my goal.

[6] Someone would prefer another term, say "confidence", but this is just a (reasonable) technical convention, not the real use and meaning of such words.

A very important case of goal-adoption (relevant for trust theory) is "goal–adhesion", where X wants and expects that Y will adopt her goal, communicates (implicitly or explicitly) this expectation or request to Y; Y knows that X has such an expectation and adopts X's goal not unilaterally and spontaneously, but also because X wants so. Thus not only Y adopts X's goal *p*, but he also adopts X's goal that Y adopts her goal *p*. In social trust frequently Y's adoption (cooperation) is precisely due to X's expectation and trust in Y's adoption; and X relies on this response and adhesion.

We agree with Hardin (2002) that there is a restrict notion of social trust based on *the expectation of adoption* (or even *adhesion*), not just on the prediction of a favorable behavior of Y. When X trusts Y in a strict social sense and counts on him, she expects that Y will *adopt* her goal and this goal will prevail—in case of conflict with other active goals. That is, X not only expects an *adoptive goal* by Y but an *adoptive decision and intention*. A simple *regularity* based prediction or an expectation simply based on some role or norm prescribing some behavior to Y, are not enough—we agree with Hardin—for characterizing what he calls "trust in strong sense", the "central nature of trust"; what we call "genuine social trust".

However, in our view, Hardin is not able to consider the broad theory of goal-adoption, and—given his notion of "encapsulated interests—provides us with a restricted and reductive view of it, just based on self-interest (wrongly meaning "selfishness"). The various authors searching for a socially focused and more strict notion of trust go in this direction, but using non general and non well defined notions, such as: *benevolence, good-will, other-regarding attitude, benignity* (Hart 1988, p. 188), *altruism, social-preferences, reciprocity*.

The fact that a genuine social trust is based/relies on Y's adoption should not be misinterpreted. One should not confuse "goal-adoption" with *specific motives* for adopting. Claiming that X counts on Y's adoptive intention is not to claim that she counts on Y's altruism, benevolence, good will, social preferences, respect, reciprocity, or moral norms. These are just specific sub-cases of the reasons and motives why Y is supposed to adopt X's goal (to act for her). X might count on Y's willingness to be well reputed (for future exchanges), or on his desire to receive gratitude or approval, or to avoid blame or sanctions, or on his self-approval, etc. Y can be fully self-interested.

To realize this it is necessary to keep in mind that the usual structures of goals are means-end chains: not all goals are "final goals"; they can be *instrumental* goals, simple means for higher goals. Thus, on the top of an adoptive and adopted goal there can be other goals that *motivate* the goal-adoption. For example, I can do something *for* you, just in order to receive what I want for me, what you promised to me.

Adam Smith (1776, Book 1, Chap. II): "It is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages." However, when I order the brewer to send me a box of beer and I send the money, I "trust" he will give me the beer.

There are three kinds of social "goal-adoption" (Conte and Castelfranchi 1995):

- *instrumental*: for my own returning profit, like in exchange, but not only in exchange. I adopt a goal of yours (help you) because this is convenient for some higher personal and selfish goal of mine.
- *cooperative* (in strict sense), where X and Y have one and the same aim, and depend on each other ("mutual dependence"). One might consider this kind of adoption just as a special case of "instrumental" adoption. In a sense this is correct; however, it really deserves to be distinguished from the previous one. They make in fact opposite predictions. While in *instrumental adoption* a rational agent should act as a cheater, should defeat the other (if X has obtained the beer he has no reason to satisfy the expectation of the brewer), this is not true in strict cooperation. If X cheats Y she cheats herself. Defeating is self-defeating. In fact, since they have the same goal and depend on each other (that is, the actions of both of them are necessary for the realization of the goal), if X will not do her share she will not achieve the goal.
- *terminal* (that is non-instrumental, a goal per sè, in itself). This is the case of true altruism (if it exists) either due to emotional impulses, and to affective relationships and attitudes (parents-children; pity and sincere charity; friendship, etc.); or due to a "cold" generosity (priority to the welfare of the other); or to moral terminal values and principles. The aim, goal, motive of X is the good, the welfare of Y; that is the satisfaction of Y's important goals. And this is a *final* goal.

X can trust Y, and trusts that Y will act as expected, *for any kind of adoption*, also (or better, usually) for the "instrumental" adoption (with both external or internal incentives); and the fact that Y will adopt X's goal doesn't necessarily mean that he is benevolent or good-willing or altruistic or moral towards X. Trust in Y doesn't presuppose that Y is "*generous*" or that he will make "*sacrifices*" for X; he can strictly be selfish.

Y can perfectly be self-motivated or interested (autonomous, guided by his own goals) and can even be selfish or egoistic; what matters is that the intention to adopt X's goal (and thus the adopted goal and the consequent intention to do *A*) will prevail on other non-adoptive, private (and perhaps selfish) goals of Y. But this only means that:

- Y's (selfish) motives for adopting X's goal will prevail on Y's (selfish) motives for not doing so and giving precedence to other goals.

So, X can count on Y's doing as expected, in X's interest (and perhaps in Y's interest).

Trustworthiness is a social "virtue" but not necessarily an altruistic one. This makes also clear that not all "genuine" trust is "normative" (based on norms) (Jones A. 2002; Baier 1986) (for example, the generous impulse of helping somebody who is in serious danger is not motivated by the respect of a moral/social norm, even if this behavior (later) is socially/morally approved).

A very important notion in goal-adoption is the notion of "concern". How much the goal of X is important for Y; how much Y is concerned with/by X's interest.

That is, what is for Y the *value* of X's goal *Gx*, or better of X achieving her goal. This value is determined by:

(1) the reasons (higher motivations) that Y has for adopting X's goal, and their value for him;
(2) the value of *Gx* for X as perceived by Y.

It is precisely on this basis that the adopted goal will prevail or not against possible costs, against other private conflicting goals of Y, and thus will possibly become/produce an adoptive *intention* of Y; and it will also—as intention—persist against possible new interferences and temptations.

It is precisely on Y's *concern* for her goal (not to be confused with benevolence, good will, benignity,…) that X relies while betting on Y's adoptive intention and persistence. She also has some "theory" about the reasons and motives why Y should be concerned with her welfare and adopt her goal.

4 Trust game: a Procuste's bed for trust theory

After this schematic presentation of our socio-cognitive theory of trust it is perhaps clearer why we criticize game-theoretical reductionism on trust.

Consider, for example, Pelligra's claim (2006) that "To isolate the basic elements involved in a trusting interaction we may use the Trust Game". We argue, on the contrary, that the Trust Game (as the great majority of game theoretic approaches and considerations about trust) gives us a biased and limited view of trust. It is like a Procuste's bed for the theory of trust.

Actually, the first two conditions identified by Pelligra are rather good:

(i) "potential positive consequences for the trustor" from the trustee's behavior;
(ii) "potential negative consequences for the trustor" from the trustee's behavior;

This means that X—as for her "welfare", rewards, goal-achievement—*depends on* Y; she makes herself "vulnerable" to Y, and Y gets some "power over" X.

However, one should make explicit—as for condition (i)—the fact that X *expects* (knows and wishes) such consequences; and decides to "count on" Y to realize them.

Condition (i) is only vague, insufficiently characterized (X might completely ignore that Y's behavior can produce good outcomes for her), but condition (iii) is definitely too restrictive for a general definition of the "basic elements involved in a trusting interaction":

(iii) "*temptation for the trustee or risk of opportunism for the trustor*".

This is a too restrictive prototype for trust-based interaction.

While relying and counting on Y (trusting him), X is exposing herself to risks: risks of failure, of non-realization of the goal for which she is counting on Y; and also risks of harms and damages due to her lack of diffidence and vigilance towards Y. This is a well-recognized aspect of trust. However, these risks (let us focus *in primis* on failure; the non realization of the expected and "delegated" action) are

not necessarily due to Y's "temptation". This seems true even in the classical definition adopted in Gambetta's book—inspired also by Deutsch (1973)—and accepted by the great majority of the authors in economics: "Trust is the subjective probability by which an individual, A, expects that another individual, B, performs a given action on which his welfare depends". In our view, this definition is correct while stressing that trust is basically an esteem, 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 Y" where there is also the decision to rely on Y; and it doesn't explain what is such an evaluation made of and based on: the *subjective probability* assembles too many important parameters and beliefs, that are very relevant in social reasoning. But even this definition doesn't restrict the risk implied in trust (the probability of failure), just to Y's opportunism or temptation.

On the one side, trust is also a belief (and a bet) on Y's competence, ability, intelligence, etc. (Y's trustworthiness). X might be wrong about this, and can be disappointed because of this. Y can just be unable or incompetent, and provide a very bad performance (service or product); Y can misunderstand X's request, expectation, or goals, and thus do something wrong or bad; Y can be absent-minded or forgetful, and just disappoint and damage X for this reason.

On the other side, not necessarily when X trusts that Y will do a given action and is waiting for and counting on it, Y is aware of this. There are acts of trust not based on Y's agreement or even awareness. Y can change his mind without any opportunism towards X; in such cases X's reading Y's mind and her prediction were wrong. Even if Y knows that X is relying on his behavior, he has no commitment at all towards X (especially if this is not common knowledge); he can change his mind as he likes. Even when there is a commitment and an explicit reliance (like in an exchange), not necessarily Y changes his mind (and behavior)—violating X's expectations—just for selfish opportunism. He can change his mind revising his intentions precisely in X's interest, even for altruism.

In fact, in another paper Pelligra recognizes and criticizes the fact that "most studies [in economics and GT] consider trust merely as an expectation of reciprocal behavior" while this is "a very specific definition of trust" (2006, p 6).

However, Pelligra—as we saw—in his turn proposes a very interesting but rather restricted definition, which fits Trust Game and the above-mentioned conditions (especially (iii)). He defines trust as characterized by the fact that X counts on Y's *responsiveness* to X's act of trusting ["The responsive nature of trust" is the title of his article (Pelligra 2005)]. This is too strong.

When X trusts Y—even in case of agreement—she can rely on Y's behavior not because Y will respond to her act of trusting her, but for many other reasons. X can count on the fact that there are norms and authorities (and (moral) sanctions) prescribing that behavior, independently of X's trust, and X assumes that Y is a worrying or respectful person. There might be a previous norm (independent of the fact that X is trusting Y), and X forecasts Y's behavior and is sure that Y will do as expected, just because the norm exists and Y knows it (Jones A. 2002). The fact that

X is trusting Y is *not* (in X's expectation) the reason inducing Y to a correct behavior.

However, let us assume that one wants to put aside, from a "true"/"strict" notion of trust, any kind of reason external to the interpersonal relationship (no norms, no third parties, no contracts, etc.). There is some sort of "genuine" trust (Hardin 2002; Jones 1996, 2001; Baier 1986), which is merely "interpersonal". In this perspective, one might perhaps claim that "genuine" trust is precisely based on responsiveness. But also this vision looks too strong and narrow. It might be the case that Y behaves as expected not because X trusts him but because X is dependent on him, for example for pity and help. He would do the same even if X wouldn't ask or expect anything. For example, X to Y: "Please, please! Don't say John what you saw. It would be a tragedy for me"; Y to X: "OK, be quiet!"; later, Z to X: "How can you trust him!?"; X: "I trust him because he is a sensible person, he understood my situation, he was moved".

More in general, X may count upon feelings or bonds of benevolence, friendship, love: she may count on those motives for Y doing what expected; not on Y's response to X's trust in him; as in the "genuine" trust of a child towards his father.

4.1 The varieties of trust responsiveness

The nice idea that we respond to a trusting act (for example, increasing our benevolence, reliability, efficacy, etc.) is a very important claim (see also Falcone and Castelfranchi 2001); but it deserves some development.

As we have shown trust has different components and aspects, so our claim is that we respond to trust in various (even divergent) ways, since we can respond to different components or faces of the trusting act, which can elicit a variety of emotions or behaviors.

For example, one thing is to react to the appreciation, the positive evaluation implicit in a decision to trust and manifested by the act of trust; or to respond to the *kindness* of not being suspicious or diffident; or to the exhibition of respect and consideration. For example, I might feel not grateful but guilty; suffering from low self-esteem and feeling that X's evaluation is too generous and deceptive and her expectation could be betrayed.

Another thing is responding to the fact that the trustor is risking on me, is counting on me, exposing herself to be vulnerable to me.

Another thing is her manifestation of being powerless, dependent on me. This for example can elicit two opposite reactions. On the one side, the perceived lack of power and the appeal to me is the basis of possible feelings of pity, and of a helpful, benevolent disposition. On the other side, this can elicit a sense of exploitation, of profiting, which will elicit anger and refuse of help: "Clear! She knows that eventually there will be this stupid guy (me!) taking care of that! She counts on this".

We do not have a complete and explanatory theory of all possible reasons why trust elicits a behavior corresponding to the expectations.

4.2 Trusting as signaling

It is clear that in those cases where the act or attitude of trust is supposed to elicit the desired behavior, it is important that Y has to know (or at least to believe in) X's disposition. This applies in both cases: when X just trusts and expects; when X is cooperating (doing something for Y) because she trusts Y and expects a given behavior. Since X plans to elicit an adoptive behavior by Y as a specific response to her act, she must ascertain that Y realizes her act towards him and understands its intentional nature (and—in case of cooperation—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 Y and to mean all that (Schelling 1960; Camerer 1988; Castelfranchi 2004).

X's cooperation in view of some form of intentional reciprocation (of any kind) needs to be a *behavioral implicit communication act* because Y's understanding of the act is crucial for providing the right motive for reciprocating. The same is for X's reliance on Y aimed at inducing Y's adoption.

This does not mean that necessarily X intends that Y understands that she intends to communicate (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, she does something in favor of Y (say, a gift). It is not necessary (and sometimes it is even counterproductive) that Y realizes the selfish plan of X, and thus the fact that she wants him to realize that she is doing something "for" him and *intends him to recognize 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 a recognition by Y: X's intention to favor Y must be recognized, but X's intention that Y's recognizes this does not need to be recognized (Castelfranchi 2004).

As we just highlighted in Sect. 4.1 the act of trusting is an ambiguous "signal", conveying various messages, and different possible meanings. And a cognitive agent—obviously—reacts to the *meaning* of the event, which depends on his active interpretation of it.

5 Trust and reciprocity

In sum, the Trust Game is not a good general framework for trust theory, and we do not have in this literature a good general definition (theory) of trust. The notion and the theory are arbitrarily restricted. Trust responsiveness is a very fundamental phenomenon, but Pelligra does not consider all its facets. In particular he does not consider: all the possible motives and mechanisms for positive (conform) responsiveness; possible counterproductive (negative) responses; and other dimensions, relative to competence and quality, and not to Y's motivation.

Given that trust does not necessarily presuppose reciprocity, and vice versa, let's now give a look at several interesting relationships between trust and reciprocity. Let's consider some of them:

(i) *X trusts that Y will reciprocate his adoption.*

She is *betting* on Y reciprocation.

(ii) *If Y reciprocates, X will trust him 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, 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 and 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. When Y is "generous" in giving, that is, he is giving even if it is not "due", or more than "due", this produces not only gratitude but also trust for the future. The more generous Y is, the more credible is his behavior as a *sign* of his "benevolence".

(iii) *Y reciprocates so that X trusts him next time (and exchanges with him, chooses him as partner).*

That is Y is paying a cost for acquiring some trust-capital (Castelfranchi et al. 2006); 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 her.

There is some sort of "meta-trust": "I trust Y, since he 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 *for* Y to trust X; more precisely "since and in order to"; it is some sort of self-fulfilling prophecy.

There is a clear psychosocial phenomenon of trust propagation where trust creates trust while diffidence creates hostility. If X trusts 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 a possible moral norm. We believe that it is mainly due to:

- the fact that while trusting Y, X makes herself dependent and vulnerable to 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, that 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 X, at her turn, reliable, safe.

Nevertheless, we believe that such a moral 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 it is responsible of blame, shame, etc.

6 Concluding remarks

We have argued against *the idea that trust has necessarily to do with contexts that require "reciprocation"; or that trust is trust in the other's reciprocation*.

A multi-layer 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 *will*, a *decision* to rely on the others that makes us dependent on and vulnerable to them, as well as a concrete *act* of reliance based on this, and a consequent *social relation*.

We have also implicitly adopted 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)* (Cialdini 2001).

On the base of this conceptual clarification and analytic work, it has been argued that we do not necessarily trust people because they will be willing to reciprocate; and that we do not necessarily reciprocate for reciprocating. Trusting people (also in strict social situations, with mutual awareness) means to count on their "adopting" our needs, doing what we expect from them, for many possible motives (from altruism to norm-keeping, from fear of punishments to gratitude, from sexual attraction to reputation and social approval, etc.); reciprocating or obtaining reciprocation are just two of them. However, the theory of how trust elicits reciprocation and trust, and how reciprocation builds trust, is an important part of the theory of trust as personal and collective capital.

Trust has for sure an enormous importance in economies and in economics (for exchange, market and contracts, for agency, for money and finance, for organizations, for reducing negotiation costs, and so on), as well as in politics (the foundational relations between citizens and government, laws, institutions), etc. However, this concerns all kinds and dimensions of trust; not only those aspects needed in strategic games.

Acknowledgments This research has been carried out within the ESF Project "The Social and Mental Dynamics of Cooperation", and the "For Trust" Project, Agence Nationale de la Recherche, France.I would like to thank Rino Falcone, co-author of our theory of trust; Luca Tummolini and Francesca Marzo for our discussions on cognition and economics, on reciprocity, etc.; Vittorio Pelligra for an interesting

debate; and an anonymous reviewer of this journal for precious editing remarks, as well as Federica Mattei for her help.

References

Baier A (1986) Trust and antitrust. Ethics 96:231–260
Barber B (1983) The logic and limits of trust. Rutgers University Press, New Brunswick
Camerer C (1988) Gifts as economic signals and social symbols. Am J Sociol 94:S180
Castelfranchi C (1998) Modeling social action for AI agents. Artif Intell 103:157–182
Castelfranchi C (2004) Silent agents. From observation to tacit communication. Modeling other agents from observations: MOO 2004—WS at the International joint conference on autonomous agents and multi-agent systems, July 19, 2004 URL: http://www.cs.biu.ac.il/~galk/moo2004/
Castelfranchi C, Falcone R (1998) Principles of trust for MAS: cognitive anatomy, social importance, and quantification. In: Proceedings of the international conference on multi-agent systems (ICMAS"98), Paris, pp 72–79
Castelfranchi C, Falcone R (2000) 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, Electronic proceedings
Castelfranchi C, Falcone R (2008) Trust theory. Wiley, London (in press)
Castelfranchi C, Falcone R, Marzo F (2006) Being trusted in a social network: trust as relational capital. In: Proceedings of Trust 2006-4th international conference on trust management, Pisa, pp 16–26
Castelfranchi C, Lorini E (2003) Cognitive anatomy and functions of expectations. Proceedings of IJCAI"03 workshop on cognitive modeling of agents and multi-agent interactions. Acapulco
Cialdini RB (2001) Influence: science and practice, 4th edn. Allyn & Bacon, Boston
Conte R, Castelfranchi C (1995) Cognitive and social action. UCL Press, London
Deutsch M (1973) The resolution of conflict. Yale University Press, New Haven and London
Falcone R, Castelfranchi C (2001) The socio-cognitive dynamics of trust: does trust create trust? In: Falcone R, Singh M, Tan YH (eds) Trust in cyber-societies. Integrating the human and artificial perspectives. Springer, LNAI 2246, Heidelberg, pp 55–72
Gambetta D (ed) (1988) Trust: making and breaking cooperative relations. Basil Blackwell, New York
Hardin R (2002) Trust and trustworthiness. Russel Sage Foundation, New York
Hart K (1988) Kinship, contract and trust: the economic organization of migrants in an African city slum. In: Gambetta D (ed) Trust: making and breaking cooperative relations. Basil Blackwell, New York
Holton R (1994) Deciding to trust, coming to believe. Austr J Philos 72(1):63–76
Jones AJ (2002) On the concept of trust decision. Support Syst 33(3):225–232. Special issue: Formal modeling and electronic commerce
Jones K (1996) Trust as an affective attitude. Ethics 107:4–25
Jones K (2001) Trust: philosophical aspects. In: Smelser N, Bates P (eds) International encyclopedia of the social and behavioral sciences. Elsevier, Amsterdam, pp 15917–15922
Kurzban R (2003) Biological foundation of reciprocity. In: Omstrom E, Walker J (eds) Trust, reciprocity: interdisciplinary lessons from experimental research. Sage, New York, pp 105–127
Luhmann N (1979) Trust and power. Wiley, New York
Luhmann N (1990) Familiarity, confidence, trust: problems and alternatives. In: Gambetta D (ed) Trust, Chap. 6. Basil Blackwell, Oxford, pp 94–107
Mashima R, Yamagishi T, Macy M (2004) Trust and cooperation: a comparison between Americans and Japanese about in-group preference and trust behavior. Jpn J Psychol 75:308–315
Miceli M, Castelfranchi C (2002) The mind and the future: the (negative) power of expectations. Theory Psychol 12(3):335–366
Pelligra V (2005) Under trusting eyes: the responsive nature of trust. In: Sugden R, Gui B (eds) Economics and sociality: accounting for the interpersonal relations. Cambridge University Press, Cambridge
Pelligra V (2006) Trust responsiveness: on the dynamics of fiduciary interactions. Working Paper CRENoS No 15
Rousseau DM, Sitkin S, Burt RS, Camerer C (1998) Not so different after all: a cross-discipline view of trust. Acad Manage Rev 23(3):393–404
Schelling TC (1960) The strategy of conflict. Harvard University Press, Cambridge

Smith A (1776) An inquiry into the nature and causes of the wealth of nations. Edited by E Cannan E, 1904, Methuen & Co, London
Williamson OE (1985) The economic institutions of capitalism: firms, markets, relational contracting. The Free Press, New York
Williamson OE (1993) Calculativeness, trust, and economic organization. J Law Econ 36(April):453–486
Yamagishi T (2003) Cross-societal experimentation on trust: a comparison of the United States and Japan. In: Omstrom E, Walker J (eds) Trust, reciprocity: interdisciplinary lessons from experimental research, Sage, New York, pp 352–370