8

The Economic Reductionism and Trust (Ir)rationality

Trust is a traditional topic in economics, for obvious reasons: economic relationships, first of all exchange, presuppose that *X* relies on *Y* for receiving what she needs (an item or a service, work or salary); she has to trust *Y* on both competence and quality, and on his reliability (credibility, honesty, fidelity). Trust is the presupposition of banks, money, commerce, companies, agency, contracts, and so on.

So, trust has been the subject of several approaches by economists and social scientists ((Williamson, 1993), (Axelrod and Hamilton, 1981) (Yamagishi, 2003) (Pelligra, 2005; 2006]). Many of them (often out of a desire to find a simple measure, some quantification,[1]) are very reductive, both from a psychological and a social point of view; the notion/concept itself is usually restricted and sacrificed for the economic framework.

In addition, a lot of interesting considerations on trust have developed (a 'trust game' for example, (Joyce *et al.*, 1995], (Henrich *et al.*, 2004]) around game theory: a recently growing domain (Luce and Raiffa, 1957) (Axelrod and Hamilton, 1981), (Shoham and Leyton-Brown, 2009).

In this chapter we aim to discuss some of those ideas (which appear, quite diffused, in other disciplines), showing why they are too reductive and how they miss fundamental aspects of the trust phenomenon, crucial even in economics.

We will discuss the particular (and for us not acceptable) formulation of trust by Deutsch (Deutsch, 1985) and in general the relationship between trust and irrationality (of the trustee or of the trustor); the very diffused reduction of trust to subjective probability; Williamson's eliminativistic position (Williamson, 1993); trust defined in terms of risk due to *Y*'s temptation and opportunism; and as an irrational move in a strategic game; the reductive notion of trust in the trust game, and in some socio-economic work (Yamagishi, 2003)); why (the *act of*) *trust* cannot be mixed up and identified with the *act of cooperating* (in strategic terms); the wrong foundational link between trust and reciprocity.[2]

[1] We call this attitude: *'quanificatio precox'*.

[2] See for example (Pelligra, 2005) and (Castelfranchi, 2009).

Trust Theory: A Socio-Cognitive and Computational Model Cristiano Castelfranchi and Rino Falcone
 ISBN: 978-0-470-02875-9

8.1 Irrational Basis for Trust?

8.1.1 Is Trust a Belief in the Other's Irrationality?

The predictability/reliability aspect of trust has been an object of study in social sciences, and they correctly stress the relationship between sincerity, honesty (reputation), friendliness, and trust. However, sometimes this has not been formulated in a very linear way; especially under the perspective of game theory and within the framework (the mental syndrome) of the *Prisoner Dilemma*.

Consider for example the notion of trust as used and paraphrased in Gambetta's interdisciplinary book on trust, on the basis of (Deutsch, 1958). It can be enounced like this: 'When I say that I trust Y, I mean that *I believe that, put on test, Y would act in a way favorable to me, even though this choice would not be the most convenient for him AT THAT MOMENT'*.[3]

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

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

As usual in arguments and models inspired by rational decision theory or game theory, together with 'rationality', 'selfishness' and 'economic motives' (utility, profit) are also smuggled. We disagree about this reading of rationality: when I trust Y in strong delegation (social commitment by Y) I'm not assuming that he – by not defeating me – will act irrationally, i.e. against his interests. Perhaps he acts 'economically irrationally' (i.e. sacrificing in the meanwhile his *economic* goals); perhaps he acts in an unselfish way, preferring some altruistic or pro-social or normative motive to his selfish goals; but he is not irrational because he is just following his subjective preferences and motives, and those include friendship, or love, or norms, or honesty, avoiding possible negative feelings (guilt, shame, regret, ...), etc. It is not such a complicated view to include within the subjective motives (rewards and outcomes) the real spectrum of human motives, with their value, and thus determining our choices, beyond the strictly (sometimes only with a short life) 'economic' outcomes.

Thus when X trusts Y she is simply assuming that other motivations (his values) will in any case prevail over his economic interests or other selfish goals. So we would like to change

[3] See also recently (Deutsch, 1958): '.... *a more limited meaning of the term implies that the trustworthy person is aware of being trusted and that he is somehow bound by the trust which is invested in him. For this more specific meaning, we shall employ the term 'responsible.' Being responsible to the trust of another implies that the responsible person will produce 'X' (the behavior expected of him by the trusting individual), even if producing MY' (behavior which violates the trust) is more immediately advantageous to him.'*. See also (Bacharach and Gambetta, 2001) '*We say that a person 'trusts someone to do X' if she acts on the expectation that he will do X when both know that two conditions obtain:*

– *if he fails to do X she would have done better to act otherwise, and*

– *her acting in the way she does gives him a selfish reason not to do X.'*. Notice in this definition the important presence of: 'to do', 'acts', 'expectation'.

the previous definition to something like this: 'When I say that I trust *Y*, I mean that *I believe that, put on test, Y would act in a way favorable to me, even though this choice would not be the most convenient for his private, selfish motives at that moment, but the adopted interests of mine will – for whatever reason and motive – prevail'*.[4]

At this level, *trust is a theory and an expectation about the kind of motivations the agent is endowed with, and about which will be the prevailing motivations in case of conflict.* This preserves the former definition (and our definition in Chapter 1) by just adding some specification about the *motives* for *Y*'s reliability: for example, the beliefs about *Y*'s morality are supports for the beliefs about *Y*'s intention and persistence. I not only believe that he will intend and persist (and then he will do), but I believe that he will persist *because of certain motives*, that are more important than other motives that would induce him to defection and betrayal. And these motives are already there: in his mind and in our agreement; I don't' have to find new incentives, to think of additional prizes or of possible punishments. If I am doing so (for example, promising or threatening) I don't really trust *Y* (yet) (see Section 9.5.2).

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

This is the crucial link between 'trusting' and the image of 'a good person'. 'Honest' is an agent who prefers his goal of not cheating and not violating norms to other goals such as pursuing his own private benefits.[5] Social trust is not only a mental 'model' of *Y*'s cognitive and practical capacities, it is also a model of his motives and preferences, and tells us a lot about his morality.

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

As we explained in Chapter 2 these beliefs about *Y*'s virtues and motives are sub-species of trust, not just beliefs and reasons supporting our trust in *Y*.

Rational Is Not Equal to Selfish

As we have just seen, it is incorrect to (implicitly) claim that the trustor relies on the trustee's irrationality (not acting out of self-interest, not pursuing their own goals and rewards); *the trustor is just ascribing to the trustee motives (perhaps unselfish ones) which will induce Y to behave conformingly to X's expectations*.

[4] Notice that this is a definition only of the most typical social notion of trust, where *X* relies on *Y*'s *adoption* of her goals (Chapter 2). As we know, we admit weaker and broader forms of trust where *Y* is not even aware of *X*'s delegation.

[5] It is important to remind us that this is not necessary and definitional for trust: I can trust *Y* for what I need, I can delegate him, *just relying on his selfish interests* in doing as expected. Like in commerce (see Adam Smith's citation in Chapter 2, Section 2.8).

However, is it irrational to count upon non 'official', non economic (monetary), non selfish, and sometimes merely hidden and internal rewards of *Y*? Is this equal to assuming and counting on the trustee's irrationality? Our answer is: Not at all!

The identification of 'rationality' with 'economic rationality' is definitely arbitrary and unacceptable. Rationality is a merely formal/procedural notion; has nothing to do with the contents (Castelfranchi, 2006]. There are no rational 'motives'. No motive can be per se irrational. It is up to the subject to have one motive or another. Economic theory cannot *prescribe* to people what are the right motives to have and to pursue; it can just prescribe *how* to choose among them, given their *subjective* value and probability. Unless Economics admits not being the science of (optimal) resource allocation and rational decisions (given our preferences or, better, motives) (Lionel Robins' view), but being the science of making money, where money is the only or dominating *valid* 'motive' of a 'rational' agent.

When *X* decides to trust *Y* (where *Y* is a psychological agent), she is *necessarily* ascribing to *Y* some *objectives*, and some *motive*, which predict *Y*'s expected behavior. These objectives are very diverse, and some of them are non-visible and even *violating the official 'game' and the public rewards*. Consider the following common sense example.

Subjective versus Objective Apples (Rewards)

X and *Y* are at a restaurant and receive two apples: one ($apple_1$) is big, mature, nice; the other one ($apple_2$) is small and not so beautiful. *Y* chooses and takes the better one: $apple_1$. *X* is manifestly a bit disappointed by that. *Y* – realizing *X*'s reaction – says: *'Excuse me, but which apple would have you chosen if you had chosen first?' X*: *'I would have taken* $apple_2$*, leaving you* $apple_1$*'*, *Y*: *'And I let you have* $apple_2$*! So why you are unhappy?'*

Why is *X* unhappy if he would have taken $apple_2$ has in fact got $apple_2$, when they are in fact one and the same apple? Actually (from the subjective and interactional perspective, not from the official, formal, and superficial one) they are two very different apples, *with different values* and providing very different rewards.

- $apple_2$, when spontaneously chosen by *X*, is $apple_2$ (that material apple with its eating value) plus:
  - a *sacrifice*, compared with the other possibility (a negative reward, but a *voluntary sacrifice* that implies the following items),
  - an *internal gratification* for being polite and kind (a positive reward), and
  - an *expectation for the recognition of this from Y*, and some gratitude or approval (a positive reward).
- $apple_2$, when left to *X* by *Y*, is – for *X* – $apple_2$ (that material apple with its eating value) plus:
  - a *sacrifice*, compared with the other possibility (a negative reward, but an *imposed sacrifice*, which is worse than a spontaneous one),
  - an *impolite act towards X*, a lack of respect from *Y*'s side (a negative reward).

Thus, in the two cases, *X is not eating the same apple*; their flavor is very different. $Apple_2$ 'has a different value in the two cases', in the first case there are fewer negative rewards and more positive rewards.

As we said, to take into account those motives and rewards in *Y*, is not seeing *Y* as 'irrational'; *Y* can be perfectly *subjectively* rational (given his real, internal rewards).

Economists and game theorists frequently enough decide from outside what are the values (to be) taken into account in a given decision or game (like to go to vote, or to persist in a previous investment), they have a 'prescriptive' attitude; and *they do not want to consider the hidden, personal, non-official values and rewards taken into account by real human subjects*. Relative to the external and official outcomes the subjects' decisions and conducts look 'irrational', but considering their real internal values, they are not at all.

Finally, it would be in its own right a form of irrational assumption, to ascribe to the other an irrational mind, while ascribing to ourselves rational capabilities; trusting the others assuming that they are different from us, and exploiting their presumed stupidity.[6] Thus, counting on those 'strange' motives (of *any* kind) seen in *Y* is neither irrational per se nor is counting on *Y*'s irrationality.

8.2 Is Trust an 'Optimistic' and Irrational Attitude and Decision?

It is fairly commonplace for the attitude and decision to trust somebody and to rely on him is *per se'* to be considered rather irrational. Is this true and necessary? Is it irrational to decide to trust, to rely on others and take risks? Is trust too 'optimistic' and a decision or expectation that is ungrounded?

8.2.1 The Rose-Tinted Glasses of Trust

Frequently trust implies optimism, and optimists are prone to trust people. What is the basis of such a relationship and how 'irrational' is optimism? Trust frequently goes beyond evidence and proof, and doesn't even search for (sufficient) evidence and proof.[7] It can be a default attitude: *'Till I do not have negative evidence, I will assume that . . .'*; it can be a personality trait, an affective disposition (see Chapter 5). Is this necessarily an 'irrational' attitude and decision from an individual's perspective?

Optimism -1

As we have anticipated in Chapter 4, one feature defining *optimism* is precisely the fact that the subject, in the case of insufficient evidence, of a lack of knowledge, assumes all the uncertain cases in favor of her desired result. Her positive expectation is not restricted to the favorable evidence, but she assumes the positive outcome as 'plausible' (not impossible) (see Figure 8.1).

On the contrary, a prudent or pessimistic attitude, maintains the positive expectation within the limits of the evidence and data, while where there is uncertainty and ignorance the subject considers the negative eventuality (see Figure 8.2).

[6] Of course, it is also possible and perfectly rational in specific cases and circumstances, and on the basis of specific evidence, to trust Y precisely because he is naïve or even stupid.

[7] It should be clear that to us this is not a necessary, definitional trait of trust, but it is quite typical. It is more definitional of "faith" or a special form of trust: the 'blind' or 'faithful' one.

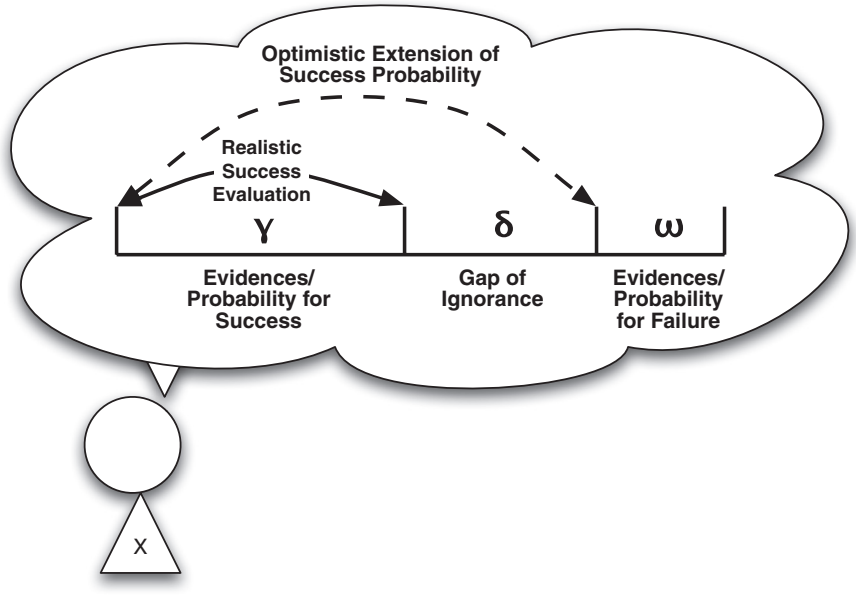

Figure 8.1 Strong Optimistic attitude

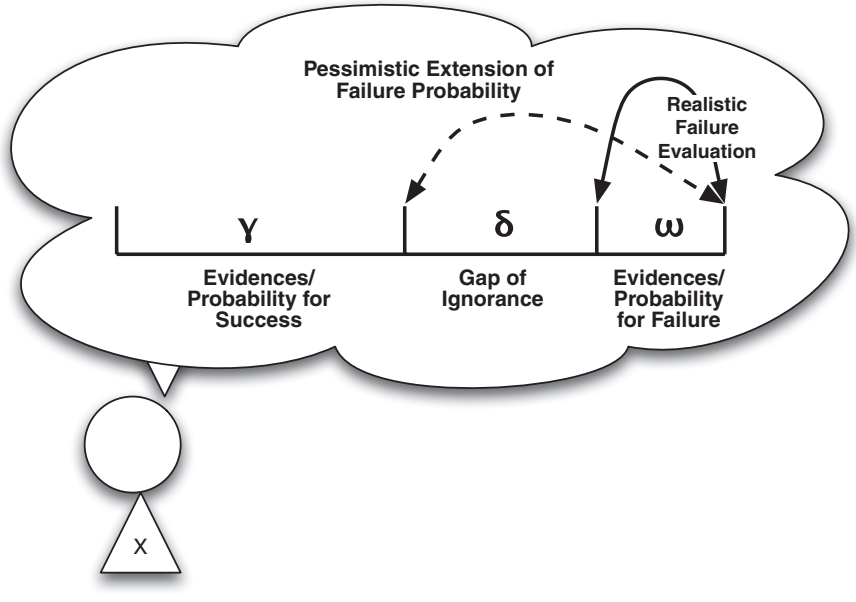

Figure 8.2 Extreme Pessimistic attitude

Extreme optimists and extreme pessimists have two opposite default assumptions and two opposite views of the 'onus of proof': does the onus lie on the prosecution side or on the defense side?

Given this (partial) characterization of optimism, when trust is not based on evidence it is an 'optimistic' attitude, since it assumes a favorable unproved outcome. Is this attitude necessarily counterproductive, too risky, or is it subjectively irrational? It seems that non-optimistic people (the more depressed ones) have a better (more realistic) perception of reality in some dimensions. For example, they have a more adequate (realistic) perception of the probability of the positive/negative events, and of their own control over events. While optimists would distort both the probability of favorable events, and the control they have on them. If something like this is true, optimism should be a disadvantage, since a distorted view of reality should lead to failures. However, a lot of studies on 'thinking positively' have shown that in several circumstances and tasks optimists fare better and that an optimistic attitude is an advantage (Scheier and Carver, 1985), (Scheier et al., 1986), (Taylor and Brown, 1988). How can we explain such an apparent contradiction: a distorted view of the world being more adaptive? We will answer this in the following paragraph.

Trust and Distrust as a Self-Fulfilling Prophecy

In sum, if the trustee knows that *X* trusts him so seriously that she is relying on him and exposing herself to risks – as we have explained – he can, for various reasons, become more reliable and even more capable, that is, more trustworthy. This is due to different possible psychological processes (Chapter 5).

The fact that *X* appreciates *Y*, has esteem for him, can make *Y* more proud and sure in his actions. He can either feel more confident, can increase his self-esteem or determination. Or *Y* can feel some sort of impulse in reciprocating this positive attitude (Cialdini, 2001). Or he can feel more responsible and obliged not to disappoint *X*'s expectation and not to betray her trust. He would bring harm to *X,* which would be unfair and unmotivated. In other words, *X*'s perceived trust in *Y* can both affect *Y*'s motivation (making him more reliable, willing, persistent,..) or his effort and competence (attention, carefulness, investments, . . .) (see also Section 8.6 on Reciprocity).

Given this peculiar dynamic of trust one may say that trust can in fact be a *self-fulfilling prophecy.* Trust in fact is – as we know– an expectation, that is, a prediction. However, this (optimistic) prediction is not just the evaluation of an *a priori* probability that is independent and indifferent to the subject evaluation.

If *X* plays the roulette game and feels sure and trustful about the fact that the next result will be 'Red', this doesn't have any affect on the probability (50%) of it being Red. But there are a lot of 'games' in human life that are not like roulette, and where an optimistic attitude can actually help.

Consider, for example, courting a woman, or competing with another guy, or preparing for an exam; in this kind of 'game' the probability of success is not predetermined and independent from the participant's attitude and expectations. Self-esteem, self-confidence, trust, etc. that is, positive expectations, make the agent more determined, more persistent, less prone to give up; and also more keen to invest in terms of effort and resources. This does change the probability of its success. Moreover, a positive, confident and self-confident attitude as perceived by the

others (*signaling*) can per se – in social interaction – be rewarding, can per se facilitate the success; for example of the courtship or of the examination.

This is why optimists are favored and more successful in many cases, compared with people with a more pessimistic and depressed attitude, although the latter might have a more realistic perception of their control (or lack of control) over events, or of the objective probabilities.

Having the 'prediction' (prophecy) is a factor that comes into the process which determines the result; thus it is a prediction that comes true, is fulfilled, thanks to its very existence. This is why trust is some sort of a 'self-fulfilling prophecy'.

However, not only can trust work like this, but also suspect and distrust can have self-realizing effects. For example, if *X* wants to be sure about *Y*, and wants to have proof and evidence about *Y*'s trustworthiness, and creates repeated occasions for monitoring and testing *Y*'s reliability (perhaps even exposing *Y* to 'temptations' in order to verify his fidelity), he is in fact altering the 'probability' of *Y*'s betrayal.

Y might be irritated by *X*'s controls and this irritation creates hostility; or *Y* can be disappointed and depressed by *X*'s lack of trust and this might decrease either his confidence or his motivation. Moreover, in general, as a trusting attitude creates a positive feeling, and also a reciprocal trust, and a diffuse trust atmosphere; analogously, diffidence elicits diffidence, suspicion or mistrust produces offence or distance.

Our explanation is as follows. In life there are two quite different kinds of 'lotteries':

a) Lotteries where the probabilities of success are *fixed a priori and independent to the subject* and their mental attitudes. This is, for example, the case of playing dice or betting at roulette. The probability is given and the subject has no influence at all on the result; it doesn't depend on their prayers, or feelings, or on how they throw the dice (excluding possible tricks), and so on. In this kind of lottery there should be more dangers for optimists and some advantage (in the case of the risk of losing opportunities) for pessimists. However, there are other – more frequent and more important – lotteries in our social life.
b) Lotteries where the probabilities of the favorable result are not given, but are influenced by the attitude, the expectations, and even the feeling of the subject. Expectations in these cases are *self-fulfilling prophecies*. This is the case, for example, of an attempt to seduce a woman; an optimistic attitude – giving to the subject and exhibiting self-confidence, providing more persistence (less perplexity) and commitment, which also might mean to the other a stronger interest and motivation – can influence the probability of success which is not a priori given. The same holds for an interview for a job or for a contract; and in many social circumstances where our success depends both on our investment and persistence (and we invest and persist proportionally to the expected probability), and on the impression we give to the partners. So, optimism might be a *real* advantage. However, this makes optimism 'adaptive', 'effective', but not yet subjectively 'rational', although it paradoxically makes 'true' and realistic an expectation, which would be ungrounded (and subjectively irrational).

Optimism may be a 'subjectively rational' attitude if we assume some sort of awareness or knowledge of its self-fulfilling mechanism. And in fact we can assume a sort of implicit belief about this: the optimist has reinforced their position by practising their 'ungrounded' belief; has acquired by experience an implicit knowledge of such an effect. Thus their belief

(expectation) is not so subjectively irrational. Unfortunately, the same reinforcement underlies a pessimistic attitude!

Optimism -2

Another fundamental feature of optimism is the fact that it focuses, makes explicit and throws attention onto the positive side of expectations. Actually expectations – including expectations on which trust is based – have a Janus nature, a double face.

In fact, since we introduce the idea of quantification of the degree of subjective certainty and reliability of *belief* about the future (the forecast), we get a hidden, strange, but nice consequence. There are other implicit opposite *beliefs* and thus *implicit expectations*.

For 'implicit' beliefs we mean in this case a belief that is not 'written', contained in any 'data base' (short term, working, or long term memory) but is only potentially known by the subject since it can be simply derived from actual beliefs. For example, while my knowledge that Buenos Aires is the capital city of Argentina is an explicit belief that I have in some memory and I have only to retrieve it, on the contrary my knowledge that Buenos Aires is not the capital city of Greece (or of Italy, or of India, and so on) is not in any memory, but can just be derived (when needed) from what I explicitly know. While it remains implicit, merely potential, until is not derived, it has *no effect* in my mind; for example, I cannot perceive possible contradictions: my mind is only potentially contradictory if I believe that *p*, I believe that *q*, and *p* implies *not q*, but I didn't derive that *not q*.

Now, a belief that '70% it is the case that *p*', logically (but not psycho-logically!) implies a belief that '30% it is the case that *not p*'.[8] This has interesting consequences on *expectations* and related emotions. The *positive expectation* that *p* entails an implicit (but sometimes even explicit and compatible) *negative expectation* (see Figure 8.3).

This means that any hope implicitly contains some fear, and that any worry implicitly preserves some hope. But also means that when one gets a 'sense of relief' because a serious threat that was expected does not arrive and the world is conforming to your desires, you also get (or can get) some exhaltation. It depends on your focus of attention and framing (Kahneman, 2000): are you focused on your worry and non existent threat, or on the unexpected achievement? Vice versa when you are satisfied about the actual expected realization of an important goal, you can also achieve some measure of relief while focusing on the implicit previous worry.

When one feels a given emotion (for example, fear), although not necessarily at the very moment of feeling it, one also feels the complementary emotion (hope) in a sort of oscillation or ambivalence and affective mixture. Only when the belief is explicitly represented and one can focus – at least for a moment – one's attention on it, can it generate the corresponding emotion.

Optimists do not think (elaborate) on the possibility of failure, the involved risks, or the negative parts of the outcome; or at least, they put them aside: they do not focus on it. In this way, they, for example, avoid the elicited feeling of worry, of avoidance, of prudence, or of non-enthusiasm *('OK, it will be good, but . . . .; not so good').*

[8] We are simplifying the argument. In fact it is possible that there is an interval of ignorance, some lack of evidence; that is that I 45% evaluate that p and 30% that *Not p*, having a gap of 25% neither in favor of *p* nor of *Not p* [29] [30].

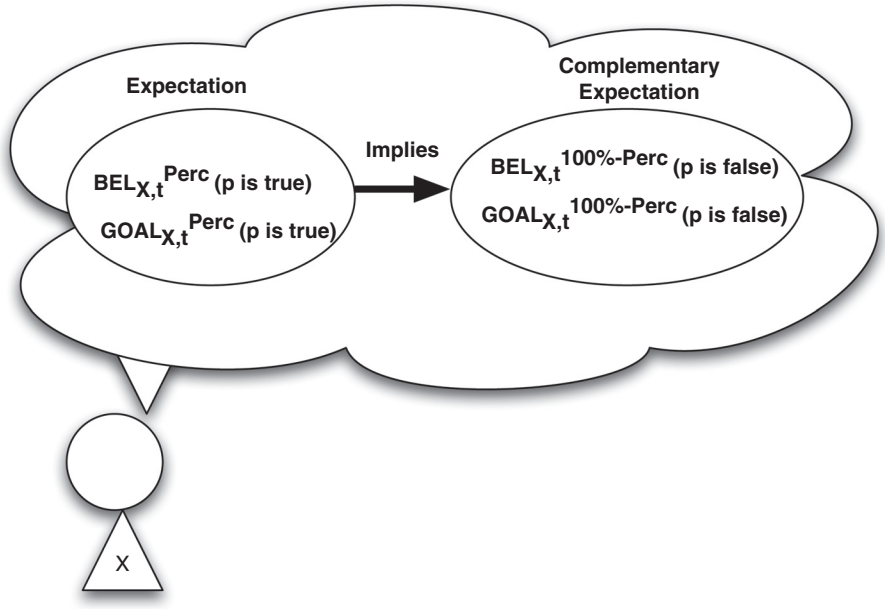

Figure 8.3 The expectation of an event (with a defined probability) implies also the expectation (with a correlated probability) of the denied event

Optimism -3

Moreover, 'optimism' in a sense is not only a prospective attitude, it is a more global attitude ('rose-tinted glasses'), even in an *a posteriori* evaluation of present results or of the past. Hence the saying: given a glass of water half-full, the optimist will see it as 'half full', while the pessimist will see it as 'half empty'. That is, the optimist will focus on the positive result, on what one has got, on the (even small) achievement or gain; and she will balance her disappointment with such a consolation. The pessimist, on the contrary, will focus on what has not been achieved, on what is lacking, and this will destroy, make marginal what has been obtained.

It is important to notice how these different perceptions of reality will reinforce the subjective prospective attitudes: the optimist will have their view confirmed that the positive expectation was not so wrong and that it is realistic to make good predictions; their attitude will be reinforced. The same for the pessimistic attitude! Both attitudes are 'self-fulfilling prophecies', self-realizing predictions; thus, both of them produce the expected results (or see the results as expected) and reinforce and reproduce themselves (Castelfranchi, 2000).

Is trust a form of optimism? An optimistic attitude? No; not necessarily. The fact is that without trust there is no society (Locke, Parsons, Garfinkel, Luhmann, etc.; see Chapter 9): so both optimistic people and pessimistic people have to trust in some way and under subjective limits and decisions. It is true: trust may be based on a very prudent and pessimistic attitude: 'I have really 'decided' – after a serious evaluation of the risks – on the basis of data and titles,

to trust this medical doctor, but I continue to be anxious and to focus on risks and worries, and to be very pessimistic about the results'.

However, *optimists are obviously more trustful and more open to trust*. This is well predicted and explained in our model. They are all for the 'non-impossible' (plausible) eventualities; so they have better expectations; they perceive less risk or do not focus on risks and thus have a lower acceptance/trust threshold; they ascribe to people a pro-social attitude, common values, non-hostility; they appreciate the gains, what has been obtained, and do not focus on what they 'didn't achieve', so they are positively confirmed in their optimism and trust-decision.

When Trust Is Irrational

As we have seen in Section 3.7 it is possible to distinguish between rational and irrational trust. One particularly interesting case, also linked with the optimism concept, is the following one, where we really have an irrational form of trust; even 'subjectively' irrational. Let's call 'anti-rational' trust a decision to trust while subjectively going against our evidences. This is not a 'spontaneous' outcome of a trust evaluation and choice; this requires an act of will; it is a decision against our evaluation. This is when we say: *'I have decided to trust him anyway, although I know that . . ., although I'm sure he will betray me'*.

This decision and attitude is quite remarkable because it violates the usual relation between trust as attitude-evaluation (*e-T*) and trust as decision (*d-T*). The usual (and rational) relation is that: since I *e-trust Y enough*, I (decide to) *d-trust Y*. And if I *d-T Y* this means that I *e-T Y*. On the contrary, when I take that attitude and force myself to trust *Y*, I do not really trust *Y*. That is, I *d-trust* him, without really *e-trusting* him. These 'subjectively' not so rational trust attitudes are not necessarily 'objectively' irrational and dysfunctional. As we said, optimism can in fact be a self-fulfilling prophecy; it can influence the chances of success.

For example, as we said, even a decision to trust against evidence and beyond the current level of evaluation and expectation, can be objectively rational just because *X*'s act of trusting *Y* can influence *Y*'s trustworthiness, can increase the chances of success: Trust creates trust (see Chapter 6).

If *X* takes into account such an influence of her attitude and decision on *Y*, and predicts and calculates this effect (see Section 6.3.3. on the dynamic of trust) her decision will also become 'subjectively' rational, since her degree of e-trust and her certainty in expectation has been increased. Moreover, the 'decision' to trust can 'locally' be irrational, but it could be part of a more general or overall decision. In saying this, *X* should have some additional reason (goal) beyond the delegated one. For example, *X* wants to publicly demonstrate courage, or persuade someone to risk and invest (be a model), etc., and this higher goal also increases the need to cover and compensate the (probable) loss on τ. So at the global level the risky trust act is part of a rational decision.

Trust Versus Faith

As we have repeatedly explained, we accept the idea that, sometimes, trust is subjectively irrational, not based on justified evidences, not grounded (or going beyond evidences), or that it is just feeling-based and intuitive (but, perhaps, grounded on analogy and experience). However, we deny that trust is necessarily and by definition only this. There is also a trust

based on evidence (observations, reports, reasoning, or learning). Trust is not contradictory to reasoning, argumentation, proof, demonstrations. In some sense some scholars mix up trust with faith.

Faith in the strict sense is not simply to believe something or *in* something without sufficient evidences. It is to believe *not* 'on the base of', 'in force of' evidence (even if evidence was there). Thus, in a sense it is renouncing evidence, refusing it; believing on a non-rational (reasons-based) ground. Frequently this is a meta-attitude; an aim. Faith rejects the need for evidence because the fact of desiring or searching for evidence or the attempt to ground our attitude on evidence is per se a 'sign' of doubt, a signal that we are doubting or might doubt. But real faith does not admit any doubt: either to doubt is prohibited ('dogma'); or we want to avoid or reject any doubt (which – notice – is some sort of meta-cognitive perception and activity).

The doubt invalidates the faith; it proves that the faith is no (longer) there. Trust on the contrary is evidence-based; not only is it not 'incompatible' with evidence (like faith), but it is inspired by evidence, signs, or experience. Not in the sense that it has always good and sufficient evidence, or that that evidence is always real 'reasons'. Actually trust can be based on different kinds of 'evidence', including feelings and emotions, intuition, practice, or mere plausibility ('not impossible'). But, in the sense that it is not *aimed at* being indifferent to evidence; evidence is very 'relevant' for trust, but not for real faith. Given this 'irrelevance' of evidence, not only is faith 'optimistic' and would also consider the 'plausible', but, it will even go against counter-evidence. It is indifferent to proof. Even if there was proof against what I believe, it is irrelevant to me, not taken into consideration. Faith aims at being non-rational (Occam).

We talk about 'faith' in a weak/broad sense, or more often 'faithful' trust, trust not justified or supported by the subjective evidences; blind; or trust not searching for evidence on the basis of some sort of meta-trust or default attitude. However, this is not the real, deep meaning of 'faith', and it is not the authentic, typical form of trust (as some authors claim).

8.2.2 Risk Perception

When we say that trust always implies some risk; that there is 'trust' and it is needed precisely because one has to assume some risk, we do not mean that this risk is necessarily explicit or focused in the mind of the trusting agent. Notice that this is not an *ad hoc* solution for an old controversial issue, it is just a general aspect of the theory of beliefs (Section 8.2.1) that we take for granted the grounding of trust attitudes (and thus decisions) on beliefs. Not only can beliefs be out of the focus of attention, or unconscious or even 'removed' in a psychoanalytic sense, but they can simply be 'implicit'. As we just said, one fundamental way in which beliefs are implicit is that they are just 'potential'; they are implied by the explicit data that we believe, but that has not yet been derived, not explicitly formulated or 'written' in some file or memory.

Subjectively speaking, for example, an agent can be fully trustful, not worrying at all, just because subjectively they don't perceive any risk and don't calculate the very small eventuality of a failure or harm. Subjectively their curve of probability is tending to a limit, is flat: 90, 95, 98% is equal to 100%, although this is actually impossible and realistically irrational (no prediction can be 100% certain about the future).[9]

[9] Even dead – contrary to moralistic *'memento mori'* – subjectively speaking is not sure: 'Who knows? Perhaps they will invent some miraculous drugs and interventions'; 'Who knows? Perhaps the water of immortality really exists; or perhaps there is resurrection and eternal life'.

So risk perception varies very much from one subject and context to another; and the risk can remain completely implicit in our mind.

Even trust beliefs can be implicit, not necessarily explicit and active. Not only in by default or affect-based forms of trust (see Chapters 4 and 5), but, for example, in a routine trust attitude, when I am used to relying on *Y*, and by experience I 'know' that he is reliable, and there is nothing to worry about.

8.3 Is Trust Just the Subjective Probability of the Favorable Event?

Our main disagreement with economists is about the following issue (as we have already seen in Section 8.1): Is trust simply reducible to *subjective probability* ?

This is in fact a dominant tradition in economics, game theory, part of sociology ((Gambetta, 1988); (Coleman, 1994)), and now in artificial intelligence and electronic commerce (Brainov and Sandholm, 1999). We argue in favor of *a cognitive view of trust* as a complex structure of beliefs and goals (in particular causal attributions, evaluations and expectations), even implying that the trustor must have a 'theory of the mind' of the trustee (see Chapter 2) ((Castelfranchi and Falcone, 1998), (Falcone and Castelfranchi, 2001)).

Such a structure of beliefs determines a 'degree of trust' and an estimation of risk, and then a decision whether to rely on the other, which is also based on a personal threshold of risk acceptance/avoidance (see Chapter 3).

In this chapter we use our cognitive model of trust to argue against probability reduction and the consequent *eliminative* behavior. We agree with Williamson (Williamson, 1985) that one can/should eliminate the redundant, vague, and humanistic notion of 'trust', *if* it simply covers the use of subjective probability in decisions. But we strongly argue against both this reduction and the consequent elimination. Trust cannot be reduced to a simple and opaque index of probability because agents' decisions and behaviors depend on the specific, *qualitative* evaluations and mental components. For example, internal or external attributions of risk/success, or a differential evaluation of trustee's competence vs. willingness, make very different predictions both about trustor's decisions and possible interventions and cautions. Let us extensively discuss some arguments against the reduction of trust to perceived probability, and eliminative behavior.

8.3.1 Is Trust Only about Predictability? A Very Bad Service but a Sure One

Very frequently an economic approach – in order to reduce trust to a well known notion and metrics (probability) – just eliminates an entire side of trust: a very typical one in common sense, in practice, and even in economic exchanges and in labor relationships: *trust as an expectation about the quality of the good or service; trust as belief about the competence, experience, skills of the trustee!*

It seems that the only concern of trust is money, and whether to be sure or not of receiving/cumulating it. But actually even money has a *quality*; not only can dollars be more reliable than euros (or vice versa), but money can be broken, forged, out of circulation, or just simulated!

Consider, for example, the already (see Chapter 1) cited definition by Gambetta: *Trust is the subjective probability by which an individual, X, expects that another individual, Y, performs a given action on which its welfare depends*.

As declared, we think that this definition stresses that trust is basically an estimation, an opinion, an expectation: a belief. It is also quite remarkable that there is no reference to exchange, cooperation, mutuality, *Y*'s awareness.

However, it is also too restricted, since it just refers to one dimension of trust (predictability), while ignoring the 'competence/quality' dimension.[10]

Moreover, to express the idea of an uncertain prediction it uses the notion of 'subjective probability' and collapses trust in this notion and measure. This is quite risky since it might make the very notion of 'trust' superfluous (see below). Clearly enough, the reliability, the probability of the desired event, has nothing to do with its degree of quality, and we cannot renounce this second dimension of trust. When we trust a medical doctor we trust both his expertise, competence, skills and his taking care of us, his being reliable, trustworthy. Trust is an, at least, bi-dimensional notion (actually – as we have shown – is a multi-dimensional construct); *Agent1*: *'Why don't you trust him? He is very reliable and well disposed'; Agent2*: *'That's true; he is very willing, but is not expert in this domain, is not well prepared'*.

8.3.2 Probability Collapses Trust 'that' and 'in'

Trust is not simply the subjective probability of a favorable event: that is, trust 'that' the desired event and outcome will be realized. If trust is *'the subjective probability by which an individual, X, expects that another individual, Y, performs a given action on which its welfare depends*', this does not only mean that the expected/desired event is an action of a given individual *Y*. What it does mean is that 'we trust *Y*', 'we trust *in Y*'. There is much more than a prediction of *Y*'s action (and the desire of that action or of its result). There is something 'about *Y*'; something we think of *Y*, or we feel towards *Y*. Is this just the estimated probability of his act or of the outcome? This definition does not capture some of the crucial kernel components of the very notion of trust: why we do not just trust 'that' *Y* will do a given favorable action, but we trust 'in' *Y*, and we see *Y* as endowed with some sort of qualities or virtues: trustworthiness, competence, reliability.

8.3.3 Probability Collapses Internal and External (Attributions of) Trust

Trust cannot be reduced to a simple and opaque index of probability because internal or external attribution of risk/success makes very different predictions about both the trustor's decisions and possible interventions and cautions.

As we saw in Section 2.7.2 one should distinguish between trust 'in' someone or something that has to act and produce a given performance thanks to its *internal characteristics*, and the

[10] In Chapter 1 we have added the following criticisms to this definition: 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 such an evaluation is made of and based on: the *subjective probability* includes too many important parameters and beliefs, which are very relevant in social reasoning. It also does not make explicit the 'evaluative' character of trust.

global trust in the global event or process and its result, which is also affected by external factors like *opportunities* and *interferences*.

Trust may be said to consist of or to (either implicitly or explicitly) imply the *subjective probability* of the successful performance of a given behavior α, and it is on the basis of this subjective perception/evaluation of risk and opportunity that the agent decides to rely or not, to bet or not on Y. However, the probability index is based on, derives from those beliefs and evaluations. In other words the global, final probability of the realization of the goal g, i.e. of the successful performance of α, should be decomposed into the probability of Y performing the action well (that derives from the probability of willingness, persistence, engagement, competence: *internal attribution*) and the probability of having the appropriate conditions (opportunities and resources *external attribution*) for the performance and for its success, and of not having interferences and adversities (*external attribution*).

Why is this decomposition important? Not only to cognitively ground such a probability (which after all is 'subjective' i.e. mentally elaborated) – and this cognitive embedding is fundamental for relying, influencing, persuading, etc., but because:

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

Let's consider the first point (a). There might be different heuristics or different personalities with a different propensity to delegate or not in the case of a weak internal trust (subjective *trustworthiness*) even with the same global risk. For example, 'I completely trust him but he cannot succeed, it is too hard a task!', or 'the mission/task is not difficult, but I do not have enough trust in him'). The problem is that – given the same global expectation – one agent might decide to trust/rely in one case but not in the other, or vice versa! In fact, on those terms it is an irrational and psychological bias. But this bias might be adaptive, for example, perhaps useful for artificial agents. There could be logical and rational meta-considerations about a decision even in these apparently indistinguishable situations. Two possible examples of these meta-considerations are:

- to give trust (and then delegation) increases the experience of an agent (therefore comparing two different situations – one in which we attribute low trustworthiness to the agent and the other in which we attribute high trustworthiness to him; obviously, the same resulting probability – we have a criteria for deciding);
- the trustor can learn different things from the two possible situations; for example, with respect to the agents; or with respect to the environments.

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

We should also consider the *reciprocal influence between external and internal factors*. When *X* trusts the internal powers of *Y*, she also trusts his abilities to create positive opportunities for success, to perceive and react to the external problems. Vice versa, when *X* trusts the environmental opportunities, this evaluation could change the trust she has for *Y* (*X* could think that *Y* is not able to react to specific external problems).

Environmental, situational, and infrastructural trust,[11] are aspects of external trust. It is important to stress that *when the environment and the specific circumstances are safe and reliable, less trust in Y is necessary for delegation (for example for transactions).*Vice versa, when *X* strongly trusts *Y*, his capacities, willingness and faithfulness, *X* can accept a less safe and reliable environment (with less external monitoring and authority). We account for this 'complementarity' between the internal and the external components of trust in *Y* for *g* in given circumstances and a given environment.

8.3.4 Probability Misses the Active View of Trust

Reducing trust to subjective probability means also reducing trust to its 'dispositional' nature, just to a (partial) evaluation, to a belief, a forecast or an expectation (including motivational aspects). We will miss the other fundamental notions of trust as decision and act: trust as betting and taking a risk on somebody, as 'relying' and 'counting on' them. Probability has nothing to do with this; it is only one possible basis (factor) for such a decision (and the perceived risk); but is not 'trusting somebody for . . .'. Moreover, the subjective probability says nothing about trusting or not *Y*; it depends on the threshold of the acceptable risk for *X*, and on the value of the foreseen gains and harms. Also after that decision and bet there is much more than a subjective probability: there is a 'positive expectation' and some worry.

8.3.5 Probability or Plausibility?

Moreover, to reduce trust to subjective probability to be taken into account in the 'subjective expected utility' guiding the decision to act or to 'delegate', is also reductive because we need a more sophisticated model of subjective evaluation of chances, including uncertainty and doubts, including 'plausibility' and gap of ignorance. We have just shown how important this is for the theory of trust, and for example the role of the 'plausibility' for characterizing optimism, or trust by default, or the real meaning of 'give credit' (see Chapter 4). Not only is a traditional probabilistic approach to trust reductive but it is also too elementary and obsolete.

8.3.6 Probability Reduction Exposes to Eliminative *Behavior: Against Williamson*

The traditional arrogance of economics and its attempt to colonize with its robust apparatus the social theory (political theory, theory of law, theory of organizations, theory of family, etc.[12]) coherently arrives – in the field of trust – to a 'collision' (Williamson, 1985) with the sociological view. The claim is that the notion of *trust* when applied in the economic

[11] They are claimed to be really crucial in *electronic commerce* and *computer mediated interaction*.

[12] In his section on 'Economics and the Contiguous Disciplines' ((Williamson, 1985) p. 251) Williamson himself gives example of this in law, political science, in sociology.

and organizational domain or, in general, in strategic interactions, is just a common sense, *empty term without any scientific added value*;[13] and that the traditional notions provided by transaction cost economics are more 'parsimonious' and completely sufficient to account for and explain all those situations where lay people (and sociologists) use the term 'trust' (except for very special and few personal and affective relationships[14]). The term trust is just for suggestion, for making the theory more 'user-friendly' and less cynical. It is just 'rhetoric' when applied to commerce[15] but does not explain anything about its nature which is and must be merely 'calculative' and 'cynical'.[16]

On the one hand, we should say that Williamson is pretty right: *if* trust is simply subjective probability, or if what is useful and interesting in trust is simply the (implicit) subjective probability (like in Gambetta's definition (Gambetta, 1988) and in the game-theoretic and rational decision use of trust), then the notion of trust is redundant, useless and even misleading. On the other hand, the fact is that trust is not simply this, and – more important – what (in the notion of trust) is useful in the theory of social interactions is not only subjective probability.

Not only is Williamson assuming a more prescriptive than scientific descriptive or explanatory attitude, but he is simply wrong in his elimination claims. And he is wrong even about the economic domain, which in fact is and must obviously be socially embedded. Socially embedded does not mean only – as Williamson claims – institutions, norms, culture, etc.; but also means that *the economic actors are fully social actors* and that they act in such a way also in economic transactions, i.e. with all their motives, ideas, relationships, etc. including the *trust* they have or not in their partners and in the institutions. The fact that they are unable to see what 'trust' adds to the economic analysis of risk[17] and that they consider those terms

[13] 'There is no obvious value added by describing a decision to accept a risk (. . .) as one of trust' ((Williamson, 1985), p. 265). 'Reference to trust adds nothing' ((Williamson, 1985) p. 265). 'I argue that it is *redundant* at best and can be *misleading* to use the term 'trust' to describe commercial exchange (. . .) Calculative trust is a contradiction in terms' ((Williamson, 1985) p. 256). Notice that he 'prescribes' the meaning of the word trust, and of its use in human sciences. Trust is only affective or moral; it cannot be based on evaluations and good evidence.

[14] '(. . .) trust, if obtained at all, is reserved for very special relations between family, friends, and lovers' ((Williamson, 1985), p. 273).

[15] 'I argue that it is *redundant* at best and can be *misleading* to use the term 'trust' to describe commercial exchange (. . .) Calculative trust is a contradiction in terms' ((Williamson, 1985) p. 256). '(. . .) the *rhetoric* of exchange often employs the language of promises, trust, favors, and cooperativeness. That is understandable, in that the *artful* use of language can produce deals that would be scuttled by abrasive calculation. If, however, the basic deal is shaped by objective factors, then calculation (credibility, hazard, safeguards, net benefits) is where the crucial action resides.' ((Williamson, 1985) p. 260). 'If calculative relations are best described in calculative terms, then the diffuse terms, of which trust in one, that have mixed meanings should be avoided when possible.' ((Williamson, 1985) p. 261). And this does not apply only to the economic examples but also to the apparent exception of 'the assault girl (. . .) I contend is not properly described as a condition of trust either' ((Williamson, 1985) p. 261). This example that is 'mainly explained by bounded rationality - the risk was taken because the girl did not get the calculus right or because she was not cleaver enough to devise a contrived but polite refusal on the spot - is not illuminated by appealing to trust'. ((Williamson, 1985) p. 267).

[16] 'Not only is 'calculated trust' a contradiction in term, but *user friendly terms,* of which 'trust' is one, have an additional cost. The world of commerce is reorganized in favor of the cynics, as against the innocents, when social scientists employ user-friendly language that is not descriptively accurate - since only the innocents are taken in' ((Williamson, 1985) p. 274). In other words, 'trust' terminology decorates and masks the *cynic reality* of commerce. Notice how Williamson is here quite prescriptive and neither normative nor descriptive even about the real nature of commerce and of the mental attitudes of real actors in it.

[17] Section 2 starts with 'My purpose in this and the next sections is to examine the (. . .) 'elusive notion of trust'. That will be facilitated by examining a series of examples in which *the terms trust and risk are used interchangeably*

as equivalent, simply shows how they are unable to take into account the interest and the contribution of cognitive theory.

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

From the traditional economic perspective all this is both superfluous and naive (non-scientific, rhetoric): common-sense notions. The economists do not want to admit the insufficiency of the economic theoretical apparatus and the opportunity of its cognitive completion. But they are wrong – even within the economic domain – not only because of the growing interest in economics towards a more realistic and psychologically-based model of the economic actor, but because mental representations of the economic agents and their images are, for example, precisely the topic of marketing and advertising (that we might well suppose has something to do with commerce).

8.3.7 Probability Mixes up Various Kinds of Beliefs, Evaluations, Expectations about the Trustee and Their Mind

We claim that the richness of the mental ingredients of trust cannot and should not be compressed simply in the subjective probability estimated by the actor for their decision. But why do we need an explicit account of the mental ingredients of trust (beliefs, evaluations, expectations, goals, motivations, model of the other), i.e. of the *mental background* of reliance and 'probability' and 'risk' components?

- First, *because otherwise we will neither be able to explain or to predict the agent's risk perception and decision*. Subjective probability is not a magic and arbitrary number; it is the consequence of the actor beliefs and theories about the world and the other agents. We do not arrive at a given expectation only on the basis of previous experiences or on the frequency of a series of events. We are able to make predictions based on other factors, like: analogical reasoning (based on a few examples, not on 'statistics'); other forms of reasoning like 'class-individual-class' (for example I can trust *Y* because he is a doctor and I trust

---

- *which has come to be standard practice in the social science literature* - (. . .)'. The title of section 2.1 is in fact 'Trust as Risk'. Williamson is right in the last claim. This emptying of the notion of trust is not only his own aim, it is quite traditional in sociological and game-theoretic approaches. For example in the conclusions of his famous book Gambetta says: '*When we say we trust someone or that someone is trustworthy, we implicitly mean that the probability* that he will perform an action that is beneficial or at least not detrimental to us is high enough for us to consider engaging in some form of cooperation with him' ((Gambetta, 1988) p. 217). What is dramatically not clear in this view is what 'trust' does *explicitly* mean! In fact the expression cited by Williamson (the 'elusive notion of trust') is from Gambetta. His objective is the elimination of the notion of trust from economic and social theory (it can perhaps survive in the social psychology of interpersonal relationships). 'The recent tendency for sociologists /the attack is mainly to Coleman and to Gambetta/ and economists alike to use the term 'trust' and 'risk' interchangeably is, on the arguments advanced here, ill-advised' ((Gambetta, 1988) p. 274).

doctors, even with no experience of them); a by default rule; an emotional activation; the existence of a norm and of an authority; etc.

- Second, because *without an explicit theory of the cognitive basis of trust any theory of persuasion/dissuasion, influence, signs and images for trust, deception, reputation, etc. is not 'parsimonious' but is simply empty*.

Let's suppose (referring to Williamson's example) that there is a girl walking in the park in the night with a guy; she is perceived by her father as under risk of assault. The father of that girl (Mr. Brown) is an anxious father; he also has a son from the same school as the guy *G* accompanying the girl. Will he ask his son *What is the probability that G will assault your sister?'* or *'How many times has G assaulted a girl?"* or *"... Has some student of your college assauledt a girl?'*.

We do not think so. He will ask his son what he knows *about G*, if he has an evaluation/information about *G*'s education, his character, his morality, his family, etc. He's not asking rhetorical questions or simply being polite. He is searching for some specific and factual information upon which to found his prediction/expectation about risk. Coleman (Coleman, 1994) too stresses the importance of information, but he is not able to derive from this the right theoretical consequences: a view of trust also in terms of justified cognitive evaluations and expectations. In his theory one cannot explain or predict which information is pertinent and why. For example, why is the artistic talent of *G* or the color of his car irrelevant?

Now, why those questions? Which is the *relevance* of those data/beliefs *about Y* for the prediction about a possible violence? Which is the *relationship* between *Y*'s 'virtues', 'qualities' (*trustworthiness*) and the prediction or better positive '*expectation*' about *Y*'s behavior? 'Trust' is precisely this *relationship*. Trust is not just reducible to the strength of a prediction: to the 'subjective probability' of a favorable event. Trust is not just a belief, or worst the degree/strength of any belief. It is a grounded belief strength (either rationally justified or affect-based), and not of any belief, but of a belief about the action of an(other) agent, and a component of a 'positive expectation'. Is Williamson's theory able to explain and predict this relation? In his framework subjective probability and risk are *unprincipled and ungrounded notions*. What the notion of trust (its cognitive analysis) adds to this framework is precisely the explicit theory of the ground and (more or less rational) support of the actor's expectation, i.e. the theory of a specific set of beliefs and evaluations *about G* (the trustee) and about the environmental circumstances, and possibly even of the emotional appraisal of both, such that an actor makes a given estimation of probability of success or failure, and decides whether to rely and depend on *G* or not.

Analogously, what can one do within Williamson's framework to act upon the probability (either objective or subjective)? Is there any rational and principled way? He can just touch wood or use exorcism or self-suggestion to try to modify this magic number of the predicted probability. *Why and how* should, for example, information about 'honesty' change my perceived risk and my expected probability of an action of *G*? Why and how should, for example, training, friendship, promises, a contract, norms,[18] or control, and so on, affect (increase) the probability of a given successful action and my estimation of it? It remains unexplained.

[18] How and why 'regulation can serve to *infuse* trading *confidence* (i.e. trust) into otherwise problematic trading relations' as Williamson reminds us by citing Goldberg and Zucker ((Williamson, 1985) p. 268).

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

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

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

Williamson is right and wrong. As we said, actually, we would agree with him (about the fact that one can/should eliminate the redundant, vague, and humanistic notion of 'trust'), but *if* and *only if* it simply covers the use of subjective probability in decisions. We have strongly argued against both this reduction and the consequent elimination. Since trust cannot be reduced to subjective probability, and needs a much richer and more complex model (and measure), it cannot be eliminated from economic decisions and models. Economics without an explicit notion and theory of trust cannot understand a lot of phenomena. For example: the real nature of the 'relational capital' (Chapter 10) and the importance of 'reputation'; that is, the role of the specific evaluations people have about me and why I invest in/for this, and the specific 'signals' my behavior sends out. The crucial role of trust (as positive attitude, as evaluation, as counting on you, as taking a risk, and so on) for eliciting a reciprocation attitude and behavior, for spreading trust, etc. (see later). The importance of trust, not only as subjective estimation/forecast, but as act and as social relation and link. And so on. In sum, reducing trust to subjective probability is a disservice to trust, and to economics.

8.4 Trust in Game Theory: From Opportunism to Reciprocity

Doubtless the most important tradition of studies on trust is the 'strategic' tradition, which builds upon the rational decision and Game Theories ((Luce and Raiffa, 1957) (Axelrod and Hamilton, 1981), (Shoham and Leyton-Brown, 2009)) to provide us a theory of trust in conflict resolution, diplomacy, etc. and also in commerce, agency, and in general in economy.

Let us start with our criticism of trust defined in terms of risk due to *Y*'s temptation and opportunism; and as an irrational move in a strategic game. Then we will consider why (the act of) trust cannot be mixed up and identified with the act of 'cooperating' (in strategic terms). Finally we will discuss *Trust game* as a wrong model for trust; and why trust is not only and not necessarily related to 'reciprocity'.

We will discuss two positions, one strongly relating trust and cooperation in *Prisoner's or Social Dilemma* situations, and later in the so called *Trust game*.

8.4.1 Limiting Trust to the Danger of Opportunistic Behavior

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

In this definition we recognize the *Prisoner's Dilemma syndrome* that gives an artificially limited and quite pessimistic view of social interaction. In fact, by trusting the other (in term of decision not simply of evaluation!), *X* makes herself 'vulnerable', she gives to the other the *possibility* of damaging her, but: does she give to *Y* even the *temptation*, the *convenience* to do so? It is true that the act of trust exposes *X* by giving *Y* an opportunity, but it is not necessarily *X* or the game structure that gives him a *motive*, a reason for damaging her[19] (on the contrary, in some cases to trust someone represents an opportunity for the trustee to show his competencies, abilities, willingness, etc.).

It is not necessary for trust that trusting Y makes it convenient for him to disappoint the trustor's expectation. Perhaps the trustor's trusting in *Y* gives him (the trustee) a reason and a motive for *not* disappointing the trustor's expectation; perhaps the trustor's delegation makes the expected behavior of the trustee convenient for the trustee himself; it could create an opportunity for strict cooperation over a common goal.

Trust continues to be trust independently of making it convenient or not for the trustee to disappoint the trustor.

Of course, there could be always risks and uncertainty, but not necessarily *conflict* in the trustee's relationship between selfish interest and broader or collective interests. If it was true that there was no trust in strict cooperation based on a common goal, mutual dependence, a common interest to cooperate, and a joint plan to achieve the common goal (Conte and Castelfranchi, 1995). While on the contrary there is trust in any joint plan, since the success of the trustor depends on the action of the trustee, and vice versa, and the agents are relying on each other.

This strategic view of trust is not general; it is an arbitrary and unproductive restriction. It is a reasonable objective to study trust in those specific and advanced conditions; what is unacceptable is the pretense of defining not a sub-kind of trust, but trust *in se*, the general notion of trust in such a way, without caring at all about the common use and common sense, about other psychological, sociological, philosophical studies and notions. Let us analyze this problem in depth, taking into account more recent and important positions which mix up trust with cooperative attitudes or actions, or based on the *Trust game* and strictly relating trust with *reciprocity*.

8.4.2 'To Trust' Is not 'to Cooperate'

As we saw, *Trust* is (also) a decision and an intention, but this decision/intention is not about 'doing something for the other', to helping or to 'cooperating' with him (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 *from* the other some sort of 'help' (intentional or non-intentional): an action useful for the trustor.

[19] The 'reciprocity' view (see below) actually seems to reverse this claim.

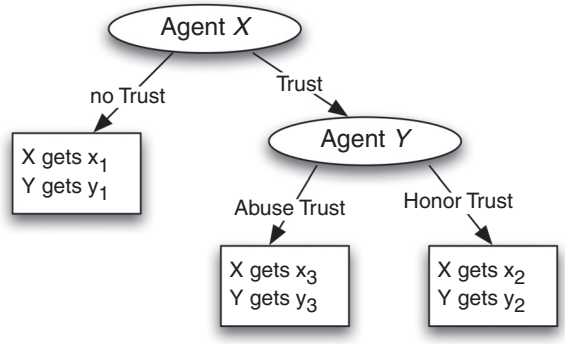

Figure 8.4 Example of Classical Trust Game

Of course, in specific 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*. An example is in fact when *X* does something for *Y* or favors *Y* while expecting some reciprocation from *Y* or for eliciting it.

This is not the only case: *X* might try to encourage an action in *Y* which would be of use to him (an action on which she decides to count and bet) not as a 'reciprocation' to 'helping' her, 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* would 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 'reciprocation' of her own 'adoption'. However, this is certainly an important group of situations, with various sub-cases which are quite different from each other from the cognitive point of view.

In some cases, *X* counts on *Y*'s feeling of gratitude, on a reciprocate motive of the affective kind. In other cases on the contrary she trusts *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 yet other cases she knows that *Y* knows the law and worries about the authority and its sanctions.[20] In these cases the act of 'cooperating' (favoring the other and risking on it) is conceived as a (partial) means 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.

For a detailed analysis of the Yamagishi approach to trust (Yamagishi & Yamagishi, 1994) (Yamagishi, 2003), and of our doubts about his mixing the two concepts of trust and cooperation, see Section 1.5.5.

8.5 Trust Game: A Procuste's Bed for Trust Theory

Figure 8.4 show the classical schema of a trust game.

[20] Notice that *X* might also adopt *Y*'s goals, while expecting *Y*'s 'cooperation', but 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 – does something for *Y*.

A trust game supposes this hypothesis: If $x_2 > x_1 > x_3$ agent X does not trust Y, make a decision with a guaranteed outcome x_1, or, trust Y and let him decide on action; If Y honors the trust, he will decide to benefit both, but if $y_3 > y_2 > y_1$ Y may abuse trust and maximize his own benefit.

Several authors use the trust game for characterizing trust. *'To isolate the basic elements involved in a trusting interaction we may use the Trust Game'* (Pelligra, 2005). On the contrary we argue 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 represents a Procuste's bed for the theory of trust.

The first two conditions for characterizing trust, as 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 it clear – as for condition (i) – the fact that X expects (knows and wishes) such consequences; and has decided 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 him), 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'*.

This is a wrong prototype for trusting interaction; too restrictive.

By 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 harm and damages due to her non-diffidence and vigilance towards Y. This is a well recognized aspect of trust (see Chapter 1). However, these risks (let us focus *in primis* on the failure; the non realization of the expected and 'delegated' action) are not necessarily due to Y's temptation (see also Section 2.8).

On the one hand, as we said, trust is also a belief (and a bet) on Y's competence, ability, intelligence, etc. X might be wrong about this, and can be disappointed because of this. Y might 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.

On the other hand, when X trusts Y to carry out a given action which she is waiting for and counting on, Y is not necessarily aware of this. There are acts of trust not based on Y's agreement or even awareness. In these cases, Y can change his mind without any opportunism towards X; it is just X's reading of Y's mind and prediction that is 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), Y changes his mind (and behavior) – violating X's expectations – just for selfish opportunism. He can change his mind, even for altruistic reasons, revising his intentions in X's interest.

8.6 Does Trust Presuppose Reciprocity?

Following the model proposed in this book, it is possible to contradict a typical unprincipled and arbitrary restriction of the notion and of the theory of trust, present in some of the 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 broader 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 to propose a restricted notion of something – fitting within a particular frame and specific issues – as the only one that is valid.

Consider, by way of an example, one of those limited kinds of definition, clearly game theory inspired, and proposed by R. Kurzban ((Kurzban, 2001), (Kurzban, 2003)): 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 to help *X* with many motives not including reciprocation; analogously, *X* can rely on *Y*'s action to have 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 not true that we either 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 would reciprocate by doing his share. This notion of trust is arbitrarily restricted and it cannot be useful to describe in detail the case where *Y* simply and unilaterally offers and promises to *X* that he will do a given action α for her, and *X* decides to count on *Y*, does not commit herself to personally perform α, and *trusts Y* for accomplishing the task. The very notion of trust must include cases like this that describe real life situations. Should we even search just for a 'behavioral' notion? *Doing nothing* and counting on others is in fact a behavior.

Even cases based on an explicit agreement do not necessarily require reciprocation. Consider a real life situation where *X* asks *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* to really do the required action. *Y* is expected to do this not out of 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 doing enough: it is not the reason *X* expects of *Y* (in fact *X* has to be grateful after the action and she is in debt); it is not what *Y* feels or the reason why he does the action; he feels that his cost greatly exceeds the received homage. Moreover, there might be other kinds of requests, based on authority, hierarchy, etc. when *X* doesn't give anything at all to *Y* 'in exchange' for the required action which is simply 'due'. But, in these cases *X* also considers *Y* to be trustworthy if she 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 diffused view is quite complicated: it mixes up a correct idea (the fact that trust – as decision and action – *implies a bet, taking some risk, being vulnerable*) 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') 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 them, 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 over the entrusted/delegated and needed action, although perhaps our costs are very limited (just a verbal request) or nothing (just exploiting an independent action and coordinating our own behavior). Sometimes, there is the risk of frustrating our goal forever 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 a good and prudent evaluator; or our social image; or other personal 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 wast our investment; risk, which is neither sufficient nor necessary.

In fact, in another article, Pelligra recognizes and criticizes the fact that *'most studies* [in economics and game theory] *consider trust merely as an expectation of reciprocal behavior'* while this is *'a very specific definition of trust'* (Pelligra, 2006).

However, Pelligra – as we saw – in his turn proposes a very interesting but rather restricted definition, which fits trust game and the previous 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 the article). We believe this is too strong.

When *X* trusts *Y* – even when agreeing on something – she can rely on *Y*'s behavior not because *Y* will respond to her act of trusting him, 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 (A. Jones, 2002). The fact that *X* is trusting *Y* is *not* (in *X*'s expectation) the reason behind *Y*'s 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), (K. Jones, 1996), (K. Jones, 2001), (Baier, 1986)) which would be merely 'interpersonal' (see Chapter 2). From this perspective, one might perhaps claim that 'genuine' trust is precisely based on responsiveness. But this vision also 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 tell 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'*.

Furthermore, in general, *X* may count upon feelings or bonds of benevolence, friendship, love: he counts on those motives that make *Y* do what is expected; not on *Y*'s responsiveness to *X*'s trust in him; like in the 'genuine' trust of a child towards his father.

8.7 The Varieties of Trust Responsiveness

As for the interesting idea that we respond to a trusting act (for example, by increasing our benevolence, reliability, efficacy, etc.) we acknowledge that this 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, diffident; or to the exhibition of respect and consideration. For example, I might not feel grateful but guilty; suffering from low self-esteem and feeling that *X*'s evaluation is too generous and misleading and her expectation could be betrayed.

I could also respond to the fact that the trustor is taking a risk on me, is counting on me, exposing her vulnerabilities to me by feeling 'responsible'. The trustor's manifestation of being powerless, dependent on me, could elicit two opposite reactions. On the one hand, 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 hand, this can elicit a sense of exploitation, of profiting, which will elicit anger and refusal 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 the possible reasons why trust elicits a behavior corresponding to the expectations.

8.8 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) *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 from *Y* as a specific response to her act, she must ascertain that *Y* realizes her act toward 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), (Cramerer, 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 doesn't mean that *X* necessarily intends that *Y* understands that she intends to communicate (Gricean meta-message): this case is possible and usual, but not inevitable. Let us suppose, for example, that *X* desires some favor from *Y* and, in order to elicit a reciprocating attitude, does something to help *Y* (say, offers 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 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 *X*'s act is certainly also aimed at such recognition by *Y*: *X*'s intention to favor

Y must be recognized, but *X*'s intention that *Y* recognizes this doesn't need to be recognized (Castelfranchi, 2004).

As we have already highlighted (see Chapter 2) 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.

8.9 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*. 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 basis of this conceptual disambiguition and of our analytic model, 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 counting on their 'adopting' our needs, doing what we expect from them, out of many possible motives (from altruism to norms 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 certainly has an enormous importance in economy and thus in economics (for exchange, market and contracts, for agency, for money and finance, for organizations, for reducing negotiation costs, and so on.), 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.

References

Alloy, L.B. and Abramson, L.Y. Judgment of contingency in depressed and nondepressed students: sadder but wiser? *Journal of Experimental Psychology: general*, 108: 441–485, 1979.

Axelrod, R. and Hamilton, W. D. (1981) The evolution of cooperation. *Science*. 211: 1390–1396.

M. Bacharach and D. Gambetta (2001) Trust as type detection, in: C. Castelfranchi, Y.-H. Tan (eds.), *Trust and Deception in Virtual Societies*, Kluwer Academic Publishing, Dordrecht, The Netherlands, pp. 1–26.

Baier, A. (1986) Trust and antitrust, *Ethics*, 96: 231–260.

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

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

Bratman, M. E. (1987) *Intentions, Plans and Practical Reason*. Harvard University Press: Cambridge, MA.

Cramerer, C. Gifts as economic signals and social symbols, *America Journal of Sociology*, 94, S180, 1988.

Castelfranchi, C. (1998) Modelling social action for AI agents. *Artificial Intelligence*, 103: 157–182, 1998.

Castelfranchi, C. (2000) Through the agents' minds: cognitive mediators of social action. In: *Mind and Society*. Torino, Rosembergh, pp. 109–140.

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. URL: http://www.cs.biu.ac.il/~galk/moo2004/

Castelfranchi, C. (2009) Trust and reciprocity: misunderstandings. RISEC (Rivista Internazionale Scienze ECo-nomiche), University Bocconi.

Castelfranchi, C. and 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.

Castelfranchi, C. and 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.

Castelfranchi, C. and Lorini, E. '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.

Castelfranchi, C., Giardini, F., Marzo, M. (2006) Relationships between rationality, human motives, and emotions. *Mind & Society*, 5: 173–197.

Castelfranchi, C., Falcone, R., and Marzo, F., Being Trusted in a Social Network: Trust as Relational Capital. Proceedings of *iTrust 2006 - 4th International Conference on Trust Management,* Pisa, May, 2006, pp. 16–26.

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

Coleman, J. S. (1994) *Foundations of Social Theory*, Harvard University Press, MA.

Conte, R. and Castelfranchi, C. (1995) *Cognitive and Social Action.* London: UCL Press.

Cramerer, C. (1988) Gifts as economic signals and social symbols, *America Journal of Sociology*, 94, S180.

Deutsch, M. (1985) *The Resolution of Conflict: Constructive and destructive processes*. New Haven, CT: Yale University Press.

Deutsch, M. (1958) Trust and suspicion. *Journal of Conflict Resolution*. 2: 265–279.

Epstein, S. Coping, ability, negative self-evaluation, and overgeneralization: experiment and theory. *Journal of Personality and Social Psychology*, 62: 826–836, 1992

Falcone, R. and Castelfranchi, C. (2001) Social trust: a cognitive approach, in *Trust and Deception in Virtual Societies* by Castelfranchi C. and Yao-Hua Tan (eds), Kluwer Academic Publishers, pp. 55–90.

Falcone, R. and Castelfranchi, C. (2001) The socio-cognitive dynamics 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*, Heidelberg, Springer, LNAI 2246, pp. 55–72.

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

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

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

Hardin, R. (2002) *Trust and Trustworthiness*, New York: Russel Sage Foundation.

Hart, K Kinship, contract and trust: the economic organization of migrants in an African city slum, in D. Gambetta, (1988) ed.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis (2004) *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* Oxford University Press.

Holton, R. Deciding to trust, coming to believe, *Australasian Journal of Philosophy*, 72 (1): 63–76, 1994,

Daniel Kahneman and Amos Tversky (eds.) (2000) *Choices, Values, and Frames*, Cambridge University Press.

Kurzban, R. (2001) 'Are experimental economics behaviorists and is behaviorism for the birds?', *Behavioral and Brain Sciences*, 24: 420–41.

Kurzban, R. (2003) Biological foundation of reciprocity. In E. Omstrom and J. Walker, eds., *Trust, Reciprocity: Interdisciplinary Lessons from Experimental Research* (pp. 105–127), NY: Sage.

Jones, A. J. (2002) On the concept of trust, *Decision Support Systems*, 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 N. Smelser and P. Bates, eds., *International Encyclopedia of the Social and Behavioral Sciences*; Amsterdam: Elsevier Science, pp. 15917–15922.

Joyce, B., Dickhaut, J., and McCabe, K. (1995) Trust, reciprocity, and social history *Games and Economic Behavior*, (10): 122–142.

Luce, R. D. and, Raiffa, H. (1957) *Games and Decisions: introduction and critical survey*, New York: Wiley.

Luhmann, N. (1979) *Trust and Power*, Wiley, New York.

Luhmann, N. (1990) Familiarity, confidence, trust: Problems and alternatives, In D. Gambetta (ed.), *Trust* (Chapter 6, pp. 94–107). Oxford: Basil Blackwell.

Mashima, R., Yamagishi, T., and Macy, M. Trust and cooperation: A comparison between Americans and Japanese about in-group preference and trust behavior. *Japanese Journal of Psychology*, 75: 308–315, 2004.

Miceli, M. and Castelfranchi, C. The mind and the future: The (negative) power of expectations, *Theory & Psychology*, 12 (3): 335–366, 2002.

Pelligra, V. (2005) Under trusting eyes: the responsive nature of trust. In R. Sugden and B. Gui, eds., *Economics and Sociality: Accounting for the Interpersonal Relations*, Cambridge: Cambridge University Press.

Pelligra, V. (2006) *Trust Responsiveness: On the Dynamics of Fiduciary Interactions*, Working Paper CRENoS, 15.

Rousseau, D. M., Burt, R. S., and Camerer, C. Not so different after all: a cross-discipline view of trust. *Journal of Academy Management Review*, 23 (3): 393–404, 1998.

Scheier, M.F. and Carver, C.S. Optimisim, coping, and health: assessment and implications of generalized outcome expectancies. *Health Psychology*, 4: 219–247, 1985.

Scheier, M.F., Weintraub, J.K. and Carver, C.S. Coping with stress: divergent strategies of optimists and pessimists. *Journal of Personality and Social Psychology*, 51: 1257–1264, 1986.

Schelling, T. C. (1960) *The Strategy of Conflict*, Harvard University Press, Cambridge, MA.

Shoham, Yoav; Leyton-Brown, Kevin (2009) *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*, New York: Cambridge University Press.

Taylor, S. E., and Brown, J. D. Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103: 193–210, 1998.

Tiger, L. (1979) *Optimism: The biology of hope*. New York: Simon & Schuster.

von Neumann, John; Morgenstern, Oskar (1944) *Theory of Games and Economic Behavior*, Princeton University Press.

Yamagishi, T. (2003) Cross-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* NY, Sage, pp. 352–370.

Weinstein, N. D. (1980) Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39: 806–820.

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

Williamson, O. E. (1993) Calculativeness, trust, and economic organization, *Journal of Law and Economics*, 36: 453–486.