9

The Glue of Society

9.1 Why Trust Is the 'Glue of Society'

In the literature (and common sense) it is now commonplace to find the slogan that trust is the glue of society, it is crucial, vital in economics (financial activity, market, management, and so on), in social cooperation, in organization, for institutional effects and acts, in groups, etc. But why? What is the real reason that trust plays such an essential role in human social life in every dimension? The answer – for us – is not (simply) the need for reducing uncertainty, for feeling and acting in a more confident way; for relying on predictions: all necessary reasons it is true, but not sufficient to understand this phenomenon and all its implications.

For us, the most fundamental reason is 'sociality' per se, the Aristotelian view of the '*zoon politikon*'. Human beings are social in a basic and objective sense: *they depend on each other*; they live thanks to each other. More precisely (Castelfranchi, 1993) (Conte, 1995) human beings are different from each other, both in their skills and resources, and in their many desires and needs (and their subjective importance).

Moreover, they live in the same 'environment', that is, they *interfere* with each other: the realization of the goals of *X* is affected by the activity of *Y* in the same environment; *Y* can create favorable or unfavorable conditions. Each agent has seriously limited powers (competences, skills, resources) and cannot achieve all his/her (potential) goals; but, by exploiting the powers of others, they can satisfy and develop their goals. By exploiting others (for example, via cooperation over common goals, or via exchange, or via domination, etc.) human beings can multiply their powers and their achievement in an unbelievable way. Also, because there are powers that no single individual possesses (*co-powers*) and cannot just be 'exchanged' or unilaterally exploited, but depends on collaboration: only a multi-agent coordinated action can produce the desired outcome. However, in order for this transformation of limits and dependence into an explosion of powers be realized, *X* not only has to exploit *Y* (and possibly vice versa) but he has to 'count on' this, to 'rely' on *Y*, to 'delegate' the achievement of his own desire to *Y*'s action; to *Y*. *This is precisely trust*. Dependence (and even awareness of dependence) without trust is nothing; it is an inaccessible resource (see Chapter 10).

Trust must be based on some experience of *Y* (or similar people), on some evaluation of *Y*'s competences and features, on some expectation, and on the decision to bet on this, to take some risk while relying on *Y*. Moreover, *X* can rely on *Y*'s understanding of this reliance, on

Trust Theory: A Socio-Cognitive and Computational Model Cristiano Castelfranchi and Rino Falcone
© 2010 John Wiley & Sons Ltd. ISBN: 978-0-470-02875-9

Y's 'adoption' of *X*'s goal and hope; on a positive (cooperative) attitude of *Y* (for whatever reason) towards *X*'s delegation. This is precisely trust; *the glue of society as the transition from passive and powerless 'dependence' to active and empowering interdependence relationships.*

As we saw, trust is not just an attitude, a passive disposition, a cognitive representation or an affective state towards another person. It is also an active and pragmatic phenomenon; because cognition (and affect) is pragmatic in general: *for* and *in* action, for realizing goals. So, trust is also an action (deciding and performing an action of betting and relying on another guy) and part of social actions: exchange, collaboration, obedience, etc.

9.2 Trust and Social Order

There is a special, substantial relationship between trust and social order; in both directions: on the one side, 'institutional', 'systemic' trust (to use sociological terms), builds upon the existence of shared rules, regularities, conventional practices, etc. and relies on this, in an automatic, non-explicit, mindless way; but, on the other side, spontaneous, informal social order (not the legal ones, with control roles and special authorities) exploits this form of trust and works thanks to it (Garfinkel, 1963). In particular, the '*stabilization*' of a given order of shared practices and common rules, creates trust (expectation and reliance about those behaviors), and this diffused and *self-confirming* trust (a self-fulfilling prophecy, as we know) stabilizes the emergent social order.

Garfinkel's theory is quite important, although partial, restricted only to some form of indirect trust, (he explicitly mentions 'trust' only a couple of time in quite a long paper), strongly inspired by Parsons and Schutz and joined with (based on) a rather strange ideological 'proclamation' against the need for psychological sub-foundations,[1] which is systematically contradicted one page later and throughout the entire paper.

The main thesis of Garfinkel is that social order and social structures of everyday life, emerge and stabilize and work thanks to our natural 'suspension' of doubts, of uncertainty, of worries; our by-default assumption is that what is coming will be normal, without surprises ('perceived normality'). We build our everyday life on such economic assumption of 'normality' and of shared, 'common' expectations about 'which game we are playing' and 'which are the well-known rules of this game'. And we react to the violation of this presupposed order and normality first of all by attempting to 'normalize' the event, to reinterpret it in another normal frame and game.[2]

Expectations about those rules and regularities, and the respect of them by the others, are *constitutive* of the game we play. *'The social structures consist of institutionalized patterns of normative culture; the stable features of the social structures as assemblies of concerted actions are guaranteed by motivated compliance with a legitimate order'* (p. 189).

Not only do the subjects suspend their possible vigilance and diffidence, but, they actively and internally 'adhere' to this order, by using those normal rules and game-definition as a '*frame*' for '*interpreting*' what is happening[3] (a rather psychological process!); and by '*accepting*' the events and the rules and the expectations themselves as 'natural', 'obvious'.

[1] 'Meaningful events [for a theory of trust] are entirely and exclusively events in a person's behavioral environment, ... Hence there is no reason to look under the skull since nothing of interest is to be found there but brains' (p. 190).

[2] Also because – we should add as cognitive scientist – for our need (mechanisms) of cognitive coherence and integration, and of social integration and coherence.

[3] 'To be clear, bridge players react to the others' actions as bridge events not just as behavioral events'.

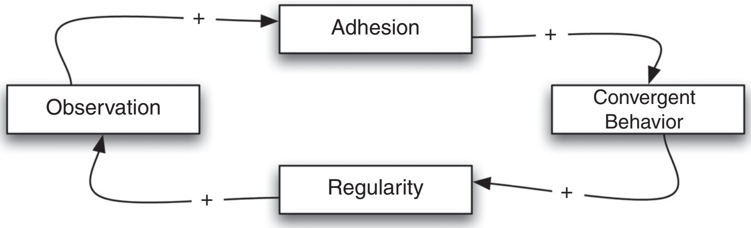

Figure 9.1 The self-reinforcing cycle of conformity

'The player ***expects*** *that the fixed reference system be binding for himself and for the others.(..) The player* ***expects*** *that his adversary has the same expectations towards him* [a rather psychological construct!]. *We define these 'constitutive expectations''.* Persons' treatments of their interpersonal environments are governed by constitutive expectancies, that is, they trust each other.

> *'The concept of trust is related to the concept of perceively normal environments as follows. To say that one person 'trusts' another means that (..) the player takes for granted the basic rules of the game as a definition of his situation, and that means of course as a definition of his relationship to others.' (p. 193)*

The events in environments are *perceived as normal* (see discussion in Section 9.6 on trust and norms).

It is interesting that in Garfinkel's analysis those expectations play cognitive, pragmatic, and normative roles; they provide the 'frame' for interpreting what is happening; they guide the decision about what to do; they create not only predictions but entitled prescriptions and commitments (Castelfranchi, 2003).

One should model in a more systematic and explicit way (also by modeling the internal mechanisms of the agents) this *cycle of routines*; how 'social structures are typically maintained in a routinary way' (Figure 9.1).

X observes a given, seemingly regular practice and on such a basis interprets it as a rule of the game, as what should be done, what to expect next time from the others; but also as what they have to do (in a given role). Thus, on one hand, they conform to that practice (also because they believe that the others expect something and prescribe this behavior); on the other hand, they create a pressure (through this behavior (*signaling*) and monitoring) on the others to conform. This actually reproduces that regularity; reinforces those beliefs and goals and makes them rule-based, routine mechanisms; confirms the validity of the 'frame' of interpretation; confirms that those are the rules of the game they are playing. Thus the cycle is self-maintaining and enforcing, and self-stabilizing (but only thanks to the *cognitive immergence* and *behavioral emergence*).

It is also important to notice and point out (Pendenza, 2000) that this kind of trust is 'natural', 'naive' (we would prefer to call it 'routine'), not based on specific reasons, reasoning, and assumptions, but just a basic reinforced attitude; just by-default, and rather automatic. It is a trust based on routines and habits. And the 'act' of trusting itself is not such a real 'decision' and 'intentional' act; it is more a functional rule-based act systematically reinforced (see Chapter 4).

9.2.1 Trust Routinization

There are various forms of trust not based on real deliberation, on intentional action, on conscious considerations of *Y*'s virtues (trustworthiness) and of possible risks (Chapter 4). Some of them are non-reflective and automatic, rule-based; due to a basic 'natural' by-default attitude, that takes for granted the reliability of *Y*, or of infrastructures, or of social rules, etc. This is the trustful disposition analyzed in Chapter 4, or Garfinkel's basic trust in spontaneous and 'natural' social order.

However, there is another form of rule-based, routine trust; it *is routinized trust*. It is when trust is initially based on careful consideration, monitoring, hesitation, a serious evaluation of *Y*'s willingness and competence. Consider, for example, a trapeze artist, starting her job with a new partner. She literally 'puts her life in the other's hands', and is very aware of his strength, attention to detail, ability, etc. But, after some months of perfect exercises and successes, she will fly in the air towards *Y*'s hands, concentrating on her own acts, and automatically relying on *Y*'s support. This holds for any "familiarization". Analogously, a blind man, who for the first time has a guide-dog and has to cross the street by following it, is very perplexed, careful and with a high perception of the risk; deciding over time to 'trust' his dog. But after they have been together for a long time, he will stop trying to control the dog all the time, and will follow it in a confident way. If this is the case in such risky relationships, a fortiori it is likely to be so in less dangerous stable reliance situations.

What, at the beginning, was an explicit belief, based on reasoning and observation, and a real decision to delegate and rely on; with successful repetitions, and exercises, becomes just a routine plan although involving (counting on) the actions of another agent; exactly as in a single-agent plan. What was a careful judgment has become some sort of reinforced classifier, and a feeling of safety and efficacy: a 'somatic marker' due to the repeated experience of success.

This routinized trust contains two form of trust: trust in the routine itself (see Chapter 4 on Routines implying trust), and procedural trust in *Y* implemented in a trusted routine.

This is also why a trust attitude and decision is not necessarily joined to an explicit consideration of risk. This idea can remain unformulated (just a logical consequence of a degree of certainty), implicit. Risks are there but not always psychologically present, although logically necessary (see Chapter 2).

9.3 How the Action of Trust Acquires the *Social Function* of Creating Trust

Trust (as attitude but especially as manifest action and signal) creates trust (see Chapter 6 and (Falcone and Castelfranchi, 2001)). A virtuous circle and thus in our model a 'function' can be created by the simple act of *X* trusts *Y*. In fact, in several ways this act can have effects that increase the probability of its reproduction (and spreading).

Since we define the 'function' of an action or of a feature an effect of it that is responsible for its reproduction without being intended, we consider these non accidental and sporadic effects of the action of trusting *Y*, which are responsible for the fact that it will be reproduced, to be its social 'function' (Castelfranchi, 2000).

There are different reasons and mechanisms responsible for this positive feedback; let's consider some of them.

a) It is true that generally speaking (but see also Chapter 6) if *X*'s act of trusting somebody (*Y*) and relying on him is successful it will increase *X*'s trust in *Y* (and also *X*'s generalized trust and a trustful attitude). Assuming that there is a reasonable distribution of trustworthy agents, and that *X* bases his decision on some reasonable criteria or experience, there is a reasonable probability that the decision to trust brings some success.

   In any case, when it brings success it will be reinforced, the probability of choosing it again will increase. While the decision not to try, not to risk cannot go in the same direction.[4] If you do not bet you can never win; if you bet you can either lose or win; if you lose (let's suppose) you will not bet again, thus being in the same situation as before.
b) The decision to trust *Y* – when known by *Y* (and frequently the act is in fact also an implicit message to *Y* 'I trust you') – may increase *Y*'s trustworthiness (see Chapter 6); either, by increasing his commitment, attention, and effort, or by reinforcing his loyalty. This will increase the chance of a good result, and thus the probability of trusting *Y* again.
c) The fact that *X* trusts *Y* (has a good evaluation of him; is not diffident towards him, decides to make herself vulnerable to *Y*), can create in *Y* an analogous non-hostile disposition, a good-will towards *X*. *Y* will have reasonable trust in *X* in return (*trust reciprocation*); but this attitude and behavior will increase the probability that *X* trusts *Y* again.
d) The act of *X* of trusting *Y* can be observed by others, can be a *signal* for them:
   - that *Y* is trustworthy (and increases the probability that they will rely on him too), or
   - that *Y* is from a trustworthy group or role, or
   - that this is a trustworthy context or community, where one can rely on the other without risk.

This will spread around trustful behaviors, that will be also perceived by *X* herself, and encourage again her trustful attitude.

Thus, in several independent ways the act of *X* trusting *Y* is responsible for its effects of increasing trust and thus of increasing the probability of its reproduction.

People are not usually aware of these effects or do not intend them, but via these effects in fact a trustful behavior has the *function* of creating trust, by spreading and reinforcing it. It is a virtuous circle, a loop (Figure 9.2).

Analogously, as trust creates and reproduces trust, distrust and diffidence enhance hostility and non-reliance and non-cooperation (*diffidence reciprocation and spreading*). This is a vicious circle in social life; what we call a 'kako-function' (Castelfranchi, 2000b). In fact it is a paradoxical 'function', since this behavior is also maintained and reproduced by its unintended effects.

[4] Except when transitory and based on the expectation of a better opportunity.

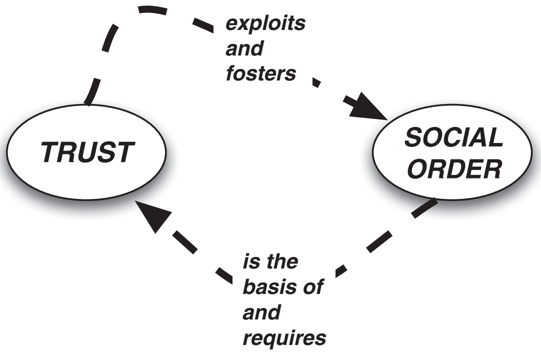

Figure 9.2 The virtuous loop between Trust and Social Order

9.4 From Micro to Macro: a Web of Trust

Let us also argue how the cognitive and interactive dynamics of trust produce the fact that trust-networks are real dynamic webs, with their own emergent macro-dynamics. They are not static, topological structures; only the sum of local relationships. Local events/changes have trans-local repercussions; the entire network can be changed; the diffused trust relationships can collapse, etc. As Annette Baier said ((Baier, 1994) p. 149): 'Trust comes in webs, not in single strands; and disrupting one strand often rips apart whole webs'.

For us Bayer is too extreme: also, merely interpersonal, dyadic (and even unilateral) trust attitudes, decisions, and relations (strands) exist. However, the phenomenon pointed out by Baier is important and must be modeled. We can also consider the individual or bilateral trust relations as very local, small and isolated webs. Actually it is true that trust attitudes and relations have a web nature.

How and why do these repercussions hold? Which are the mechanisms of this web-dynamics?

9.4.1 Local Repercussions

If *X* trusts *Y*, *W*, and *Z*, the fact that she revises her trust in *Y* can affect her trust in *W* and/or *Z*; there might be some repercussion. Not only because the trust in *Y* might have been also comparative: *X* has decided to actively trust *Y*, to choose *Y* and delegate to him, in comparison with *W* and *Z*. Thus, the success of this delegation (or the decision itself, thanks to Festinger's effect (Festinger, 1957)) can make more certain *X*'s evaluation of *Y* and change the relative strength of trust in *W* and *Z*. While the failure of that delegation can comparatively increase trust in *W* and *Z*.

There might also be other repercussions. For example some common category (see Section 6.6) or some pertinent analogy between *Y* and *Z*, and thus the success or failure of *Y* can also change the intrinsic evaluation and expectation about *Z*. The trust in *Y* is betrayed or a disaster, but *Y* and *Z* belong to the same category/role (for example, layers); *X* generalizes his disappointments and negative expectations to the whole class; thus, also *X*'s trust in *Z* collapses. And so on.

9.4.2 Trans-Local Repercussions

Also non-personal and non-local changes of the trust relationships can affect *X*'s views and the entire network. *Y* can, for example, just observe *J's* delegation to *K* and on such a basis change her own trust disposition (and decision) towards *Z*. In fact, there might be analogies between *J*'s delegation and *X*'s potential delegation, or between *K* and *Z*. Or *X* can just imitate *J*, use his example as a model or cue, not only in a pseudo-transitive way (*X*'s attitude towards *K* is derived from *J*'s attitude), but in an analogical way: if *J* trusts *K*, *X* can trust *Z* (because *X* and *J* have similar needs, and *K* and *Z* similar properties) (see again Section 6.6).

Notice that the observation of the others' behavior and their evaluation (in relation to social conventions and norms, to fairness, honesty, etc.) is a basic fundamental social disposition and ability in human beings. We do not just observe the behaviors of agents that directly concern us (exchanging or cooperating with us, competing with us, etc.); we observe agents interacting with other agents not related with us. And we in a sense provide an altruistic 'surveillance' of each other (Gintis, 1957); we evaluate them, we spread around our blame or admiration, we spread the circulating reputation. This is the most powerful instrument for social control and social order in human societies, but, it will be equally important in virtual and in artificial societies. This is exactly why 'identification' of the agents matters in the first place.

Other trans-local mechanisms for trust repercussion apart from observation of distal events, are referrals (other agents report to me their evaluation or the success/failure of their delegation (Yu and Singh, 2003), and reputation: spreading around opinions and gossip about agents (Conte and Paolucci, 2002).

Particularly delicate conditions for web effects are *default* and *generalized forms of trust*. Some of them could collapse in an impressive way (like our trust in money). If, for example, the rule of my generalized trust is:

> 'Since and until everybody trusts everybody (or everybody is not diffident and suspicious towards the others) ==> I will trust anybody (I will not be diffident)'.

This rule, given one single case of personal bad experience, or of bad observation, or referral, or reputation, can invert its valence: I become suspicious in a generalized way. And if this rule is diffused (all the agents or many of them use it) the impact will be a generalized collapse of the general trust capital and atmosphere.

Analogously, if I follow an optimistic (non prudent) default rule: 'Except I have a negative example I will assume that agents in this community are trustworthy'.

The same can hold for affective trust disposition (Chapter 5) and a generalized mood in a given context that can be wasted by just one very negative personal experience or by the contagion of different moods of others.

Of course the network is not necessarily uniform or equally connected; it might be an archipelago of non well-connected 'islands' of very connected sub-communities. Thus the propagation might just have local effects.

9.5 Trust and Contracts

> *'A contract is not sufficient by itself, but is only possible because of the regulation of contracts, which is of social origin'* E. Durkheim, (the Division of Labor in Society (1893), New York: The Free Press, 1997, p. 162).

Obviously, this *social* background includes trust, social conventions and trust in them, and in people respecting them, the authorities, the laws, the contracts.

9.5.1 Do Contracts Replace Trust?

A commonplace belief in trust theory is that we put contracts in place when and because there is no trust between the parties. Since there is no (not enough) trust, people want to be protected by the contract. The key in these cases is not trust but the ability of some authority to enforce contract application or to punish the violators. Analogously, in organizations people do not rely on trust but on authorization, permission, obligations and so forth.

As we have explained in Chapter 7 (on trust and third party relationships), for us this view is correct only if one adopts a very limited view of trust in terms of direct interaction and acquaintance, of friendliness, etc. But it is not true that 'trust', as a general category, is not there in contracts or in formal agreements and rules. In those cases (contracts, organizations) we just deal with *a more complex and specific kind of trust.* But trust is always crucial. A third party (and 'institutional') trust.

As we have said, we put a contract in place only because we believe that the agent will not violate the contract, and we count on that; and this is precisely 'trust'. We base this trust in the trustee (the belief that they will do what has been promised) either on the belief that they are a moral person and keeps their promises, or on the belief that they worry about law and punishment by the authorities (*A*). This expectation is the new level of *X*'s trust in the trustee.

As we have explained, *X* relies on a form of paradoxical trust of *Y* in *A*: *X* believes that *Y* believes that *A* is able to control, to punish, etc. Of course, normally a contract is bilateral and symmetric, thus the point of view of *Y* should be added, and his trust in *X* and in *A* when it comes to monitoring *X*. Notice that *Y*'s beliefs about *A* are precisely *Y*'s trust in the authority when they are the client, while, when *Y* is the contractor, the same beliefs are the basis of their respect/fear toward *A*.

So contracts presuppose less informal/personal trust but require some more advanced (cultural, institutional) form of trust, in Y, and in the institution.

9.5.2 Increasing Trust: from Intentions to Contracts

What we have just described are not only different kinds and different bases of trust. They can also be conceived as different levels/degrees of social trust and *additional* supports for trust. We mean that one basis does not necessary eliminate the other but can supplement it or replace it when it is not sufficient. If I do not trust your personal persistence enough I can trust you to keep your promises, and if this is not enough (or is not there) I can trust you to respect the laws or to worry about punishments.

We consider these ‘motivations’ and these ‘commitments’ not all equivalent: some are stronger or more cogent than others. As we claimed in (Castelfranchi, 1995):

> *This more cogent and normative nature of S-Commitment explains why abandoning a Joint Intention or plan, a coalition or a team is not so simple as dropping a private Intention. This is not because the dropping agent must inform her partners -behaviour that sometimes is even irrational-, but precisely because Joint Intentions, team work, coalitions (and what we will call Collective-Commitments) imply S-Commitments among the members and between the member and her group. In fact, one cannot exit a S-Commitment in the same way one can exit an I-Commitment. Consequences (and thus utilities taken into account in the decision) are quite different because in exiting S-Commitments one violates obligations, frustrate expectations and rights she created. We could not trust in teams and coalitions and cooperate with each others if the stability of reciprocal and collective Commitments was just like the stability of I-Commitments (Intentions).*

Let us analyze this point in more detail by comparing *five* scenarios of delegation:

Intention Ascription

X is weakly delegating *Y* a task τ (let’s say to raise his arm and stop the bus) on the basis of the hypothetical ascription to *Y* of an intention (he intends to stop the bus in order to take the bus).

There are two problems in this kind of situation:

- The *ascription of the intention* is just based on abduction and inferences, and to rely on this is quite risky (we can do this when the situation is very clear and very constrained by a script, like at the bus stop).
- This is just *a private intention and a personal commitment* to a given action; *Y* can change his private mind as he likes; he has no social obligations about this.

Intention Declaration

X is weakly delegating *Y* a task τ (to raise his arm and stop the bus) on the basis not only of *Y*’s situation and behavior (the current script) but also or just on the basis of a declaration of intention by *Y*. In this case both the previous problems are a bit better:

- the *ascription of the intention* is safer and more reliable (excluding deception that on the other hand would introduce normative aspects that we deserve for more advanced scenarios);
- now *Y knows that X knows about his intention* and about his declaring his intention; there is no promise and no social commitment to *X*, but at least by changing his mind *Y* should care about *X*’s evaluation of his coherence or sincerity or fickleness; thus he will be a bit more bound to his declared intention, and *X* can rely a bit more safely on it.

In other words, *X*’s degree of trust can increase because of:

- either a larger number of evidences;
- or a larger number of motives and reasons for *Y* doing τ;
- or the stronger value of the involved goals/motives of *Y*.

Promises

Promises are stronger than a simple declaration or knowledge of the intention of another agent. Promises create what we called a social commitment, which is a right producing act, determining rights for *Y* and duties/obligations for *X*. We claim that this is independent of laws, authority, punishment. It is just at the micro level, as an inter-personal, direct relation (not mediated by a third party, be it a group, an authority, etc.).

The very act of committing oneself to someone else is a 'rights-producing' act: before the S-Commitment, before the 'promise', *Y* has no rights over *X*, *Y* is not entitled (by *X*) to exact this action. After the S-Commitment such a new and crucial social relation exists: *Y* has some rights on *X*, she is entitled by the very act of Commitment on *X*'s part. So, the notion of S-Commitment is well defined only if it implies these other relations:

- *Y* is entitled (to control, to exact/require, to complain/protest);
- *X* is in debt to *Y*;
- *X* acknowledges being in debt to *Y* and *Y*'s rights.

In other words, *X* cannot protest (or even better *he is committed to not protesting*) if *Y* protests (exacts, etc.).

One should introduce a relation of 'entitlement' between *X* and *Y* meaning that *Y has the rights* of controlling α, of exacting α, of protesting (and punishing), in other words, *X* is S-Committed to *Y* to not oppose these rights of *Y* (in such a way, *X* 'acknowledges' these rights of *Y).*

If *Y* changes his mind he is disappointing *X*'s entitled expectations and frustrating *X*'s rights. He must expect and undergo *X*'s disappointment, hostility and protests. He is probably violating shared values (since he agreed about *X*'s expectations and rights) and then is exposed to internal bad feelings like shame and guilt. Probably he does not like all this. This means that *there are additional goals/motives that create incentives for persisting in the intention. X* can reasonably have more trust.

Notice also that the declaration is more constraining in promises: to lie is worst.

Promises with Witness and Oaths

Even more binding is a promise in front of a witness, or an oath (which is in front of God). In fact, there are other bad consequences in the case of violation of the promise. *Y* would jeopardize his reputation (with very bad potential consequences; see [Cas11]) receiving a bad evaluation also from the witness; or if he'd behaved badly under oath he would elicit God's punishment.

Thus if I do not trust what you say I will ask you to promise this; and if I do not trust your promise I ask you to promise in front of other people or to take an oath on it. If I worry that you might break that promise I will ask for it in writing and signed. And so on.

Contracts

Even public promises might not be enough and we may proceed by adding binds to binds in order to make *Y* more predictable and more reliable. In particular we might exploit the third party more. We can have a group, an authority able to issue norms (defending rights and creating obligations), to control violation, to punish violators. Of course this authority or reference group must be shared and acknowledged, and as we said trusted by both *X* and *Y*. Thus we have got an additional problem of trust. However, now *Y* has additional reasons for keeping his commitment, and *X*'s degree of trust is higher.

Notice that all these additional beliefs about *Y* are specific kinds or facets of trust in *Y*: *X* trusts that *Y* is respectful of norms, or that *Y* fears punishments; *X* trusts in *Y*'s honesty or shame or ambition of good reputation, etc. To state this point even more clearly: the stronger *Y*'s motive for doing τ and then the stronger his commitment to τ; the greater the number of those motives; and the stronger *X*'s beliefs about this; the stronger will be *X*'s trust in *Y* to do τ.

9.5.3 Negotiation and Pacts: Trust as Premise and Consequence

If *X* is negotiating with *Y* over a possible agreement and pact this means that she has some form and degree of trust in *Y* as (i) possible adequate and credible 'provider', and as (ii) willing and capable negotiator. She is not just wasting her time or doing this for fun. She is already betting and relying on *Y* to 'negotiate' and possibly achieve an agreement; and this is a possible bet and reliance on *Y* to do the delegated task (good, service).

Negotiation and pacts (contracts) *presuppose* some trust. However, pacts and contracts are there also to *create* some trust: the trust on which *Y* will, for example, give her money and confidently wait for a return. In fact, as we just said (Section 9.4.2), *Y*'s promise (implicit or explicit), his 'commitment', gives *X* (additional) new bases for expecting that *Y* will act as desired. *Y* now has some 'duty' to do so; there are social and moral norms binding him; and *X* is entitled to her expectations; she has some 'right' over *Y*. Thus *X* has additional trust in *Y*; she feels more sure and safe.

With a 'contract' we have the added weight of legal authorities and guaranties to protect *X* and to oblige *Y*. So, negotiation and pacts are a social mechanism for producing trust by investing trust; for building new forms and bases of trust on previous ones. However, as we said, the new layer of trust is not always 'additional'; it is also completing or complementing some possible lack and insufficiency of trust. If *X* does not trust *Y* enough on a mere informal base, she would not rely on him. Then, a promise, pact, or contract gives her sufficient trust by binding *Y*'s decision. Nevertheless, some trust must always be already there before the negotiation, or the promise, or the pact, or the contract. Why ask for a promise from *Y* if we do not believe that he is promise-sensible and bound? Why invoke the official signature of a contract if we perceive *Y* as indifferent to law, authority, sanctions, etc.?

9.6 Is Trust Based on Norms?

This is a quite diffused theory (especially in sociology; for example Garfinkel, 1963, and Giddens, 1984. See also Section 2.2.2 for our criticism on A. Jones' position).

1) On one hand, one should not over exploit the very dangerous ambiguity of the notion of 'norm' or of 'rule' in several languages, covering both a mere external descriptive 'regularity' (from Latin: 'regula', rule/norm), and a prescription or model aimed at inducing

a given conforming behavior in self-regulated systems, and the translation of this into some proximate *mechanism* affecting or producing the behavior of the system, conforming to that rule and thus 'regular'. It is not one and the same thing.

In particular it is quite different:

- saying that I trust and rely on a given predicted behavior (*based on* some perceived rule or regularity, or on some norm and conformity to it);
- saying that I trust the prediction; or that I trust the norm;
- saying that I trust the behavior in force of the explicit norm.

Is my prediction *based-on* such (perceived) regularity, on such a rule; or is that behavior *based on* that rule (affecting the mind of *Y*)? It is not at all the same thing.

2) However, even more important than this, one should be careful to preserve the very fundamental distinction made by Tommaso between*: 'Id quod intelligitur'* and *'Id quo intelligitur*': what (O) I'm thinking about, categorizing, recognizing, understanding, knowing, vs. what I'm using for thinking about O, for representing O in my mind (or externally): the representation, the scheme.[5] *I'm not thinking the representation; I'm thinking about my object of knowledge through the representation.*[6]

This clear distinction is fundamental for cognitive sciences (and semiotics).

Analogously, thanks to and through a given (implicit or explicit) 'rule' and learned regularity, I think that something *p* will happen (in the future). I do not believe – in the same way and sense – the rule (or *in* the rule). I believe *with/through/ thanks* to the rule ('Id quo'), not the rule ('Id quod').

If, for example, I believe that – since it starts raining – the ground will become wet, or if I believe that in springtime that tree in my garden will produce flowers (and I trust in this), I do *not* believe that the tree (or the rain) will follow/respect the norm. Not only do I not have some animistic, and 'intentional stance', but even less I believe that 'the rule will be respected'. I just *use* the rule (for inferring); its systematic use is a procedural, implicit assumption that it is true (reliable) and that 'it will be respected', but not an explicit belief and judgment, like my expectation about *p*.

If I strongly hope and even *trust* that she will accept my courtship this night, after my flowers, dinner, intimate atmosphere, wine, etc. as usual from my previous experiences, I do not 'trust' (believe) that 'she will respect the rule', or that 'the rule will be respected'.

Logicians seem rather insensible to this fundamental distinction between explicitly represented goals or beliefs, and merely procedural implementations. For example, one should not use the same predicate *(Bel x p)* to represent the *status/use/role* of 'being believed' of *p* in *X*'s mind, and the object 'belief'; object of various propositional attitudes: *(Goal Y (Bel X p)), (Bel*

[5] Quaestio 85; Prooemium Deinde considerandum est de modo et ordine intelligendi. Et circa hoc quaeruntur octo. Primo, utrum intellectus noster intelligat abstrahendo species a phantasmatibus. Secundo, utrum species intelligibiles abstractae a phantasmatibus, se habeant ad intellectum nostrum ut quod intelligitur, vel sicut id quo intelligitur (Thomas de Aquino, Summa Theologiae, I^{a} q. 84–89)

[6] Except I go to a meta-level, and take the representation itself (the 'significant') as my object of reflection.

Y (Bel X p)). These two '*Bel*' cannot be represented in the same way; I cannot use *(Bel X p)* to build a belief in *X*'s mind.

3) In sum, if it was true that any possible expectation, for its 'prediction' part, is based on some 'inference' and that an 'inference' is based on some 'rule' (about 'what can be derived from what'), it would be true that any trust is based on some 'rule' (but a cognitive one). However, even this is too strong; some activated expectations are not based on 'inferences' and 'rules' but just on associative reinforced links: I see *q* and this just activates, evokes the idea of *p*. This is not seriously a 'rule of inference' (like: '*If (A is greater than B) and (B is greater than C), then (A is greater than C)*'). So we would not agree that any expectation (and trust) is rule-based. However, one might expand the notion of 'rule' even to this simple and reactive 'mechanism' (mixing up the observed regularity that they produce, with a 'regula'/rule that should generate it). With such a broad and weak notion of rule, we might agree that any trust – being prediction based – is in some sense 'rule-based', it reflects some regularity and 'norm'. But not in the strict social or moral or cognitive sense; this holds only for social trust in its 'genuine' sense, based on goal-adoption and (implicit) commitments or on social norms and prescriptions.
4) Moreover, regularity is also about bad events; we also have 'negative' expectations (based on the same rules, 'norms' of any kind). Now, it is a real act of violence against the current notion of 'trust' that we are supposed to model, reducing it just to 'expectations based on perceived normality' (Garfinkel's claim).

We may have the expectation that the author of a horrible homicide will be condemned to die (given the laws of our states, and the practice of our government), both if we wish this to happen and expect it out of revenge, or if we are the killer condemned to die, or activists against the death sentence. However, if we are in favor of the death sentence, and we desire this, actually we 'trust' our authorities over this; if we are the condemned man, or the adversaries of the death sentence, we don't trust the authorities at all over this! This would be a serious distortion of the concept. This is why in our chapter about third party trust (Chapter 7) we say that this is a 'paradoxical', not true form of trust. There is a basic common mental ingredient (the belief about the future event), and this explains why the same belief becomes trust or not while just changing my role and goal. But it is not the right solution to reduce trust just to such a belief, and to call 'trust' fear and opposition.

Thus, in sum, *normality* and *regularity* are not sufficient for trust, and probably are not even necessary, if we do not extend conceptually the notion of 'rule' to cover any possible prediction device.

9.6.1 Does Trust Create Trust and does There Exist a Norm of Reciprocating Trust?

We have made it clear (Chapter 6) that it is not out of reciprocation that *Y* does the expected action after we have trusted him and decided to rely and depend on him; and also that trust is not always 'reciprocated' (even when *Y* performs the entrusted action). However, we acknowledge that there exist a property of trust to elicit trust, and we wonder about the idea that there might even exist *a norm of trust reciprocation. Since trust is not just a behavior, but a mental state* *and a feeling, it cannot really be 'prescribed', since is not really 'voluntary'*. Only the act, the intention can paradoxically be 'prescribed': 'Trust him! Rely on him!'; but not the real background disposition.[7]

However, moral (and religious) norms can impinge even on mere mental dispositions ('Do not desire . . .' 'Do not have this kind of thought'); thus there might be, and in fact it seems that there is, a social-moral *norm* about reciprocating trust: 'Since if *X* trusted you, you have to trust *X*'. To trust somebody seems to be a form of 'gentle' disposition or act, and it seems that we have to respond to a gentle act with a gentle act, to a smile with a smile.

There is a clear psychosocial phenomenon of trust propagation such that trust creates trust while diffidence creates hostility. If *X* trusts *Y*, this tends to elicit not only a 'benevolent' but also a 'trustful' attitude in *Y* towards *X*. However, we do not believe that it is mainly due to such a possible moral norm. We believe that it is mainly due to:

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

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

9.7 Trust: The Catalyst of Institutions

As we said, trust is crucial for the whole of social life (exchange, cooperation, communication, rules of conflict, etc.), however *it is in particular fundamental* (or better, foundational) *for the 'institution'* (Searle, 1995).

Together with:

- actors' recognition and assumption (acceptance) of the institutional act and effect, and with
- actors' 'as if' behavior (conforming to the assumption) (Tummolini, 2006), trust is the necessary ground on which our 'institutions' base themselves, their 'count-as' nature.

Actually, it is trust (and behavioral conformity) that 'institute' them and give them (make 'real') their artificial effects.

In fact, the social 'representation', the collective *mise en scene* (Goffman, 1959), is strictly based on compliance and complicity, on collusion; that is, on the (prescribed) assumption that

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

everybody is doing their share (starting from such an assumption). If you do not believe – or better ‘accept’ – that C is equal to D, (‘counts as’ D) it doesn’t actually count as D.

Only our cooperation and complicity creates the phenomenon, provides the needed ‘power’ and ‘virtue’ to those acts (like: signing, paying), roles (like: judges, policemen), objects (like: money, signatures). If C has to ‘count as D’ for us, then I have to pragmatically ‘count on’ its conventional effects, and thus I have to ‘count on’ you (us) for its ‘counting as D’ for you. I (have to) trust you to recognize and frame C as D, and *treat* C as D by behaving in front of C according to our convention.

Trust is not just the *glue*, but is the real *mediator*[8] *of the constructive process and mechanism* of the conventional actions/effects and the institutional building and maintenance.

9.7.1 The Radical Trust Crisis: Institutional Deconstruction

The most radical and serious economical-political crisis is in fact a *trust crisis*; when the ‘doubt’ corrodes the conventional, artificial, value, nature, and effect of *institutional* powers, actions, and objects.

I no longer believe that the court or the policeman has any authority over me (over us) since you and I do no longer recognize them. I do not believe that our money has a value, I’m not sure that the others will accept it, so why should I accept it or preserve it? I no longer believe that your act (signature, oath, declaration, etc.) has any value and effect, since to be effective as conventional-institutional act it presupposes the *acknowledgment* and *compliance* of other people, and I do not believe that the others believe in its validity and will be compliant.

This is a real institutional earthquake: I do not trust institutional authority, roles, acts, artifacts (such as money), because I do not believe that the others trust them. We move from a shared implicit trust in institutional artifacts and acts, to a *shared distrust*. But, since trust is the real foundation of their reification and effectiveness, this make them disappear: we realize, we see, that ‘the king is naked!’.

Also the political ‘representation’ is an *institutional act*. We take *X*’s (the ‘representative’) words or choices as ‘representing’ the preferences, the opinions, or at least the interests and values of his group (the people he ‘represents’); and those people believe the same and rely on this. However, if the represented people’s trust is in crisis, and thus there is no longer a real reliance and *‘delegation’* to *X*, *X* no longer represents anybody. There is a serious detachment, a crisis of the relation between people and parties, voters and deputies, which essentially is a trust-delegation-reliance relation. If I no longer believe in you or ‘count on’ you, you no longer represent me. I don’t necessarily perceive you as dishonest or selfish, but perhaps I perceive you as powerless or I perceive politics as distant and ineffective.

References

Baier, A. (1994) Trust and its vulnerabilities in *Moral Prejudices*, Cambridge, MA: Harvard University Press, pp. 130–151.

Castelfranchi, C., Cesta, A., Conte, R., Miceli, M. Foundations for interaction: the dependency theory, Lecture Notes in *Computer Science*; 728: 59–64, Springer, 1993.

[8] Trust is more than a *catalyst* for the institutional process; in fact it is also among the results of the ‘reaction’.

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

Castelfranchi, C. (2000b) Per una teoria pessimistica della mano invisibile e dell'ordine spontaneo. In Salvatore Rizzello (a cura di) *Organizzazione, informazione e conoscenza*. Saggi su F.A. von Hayek. Torino, UTET.

Castelfranchi, C. (1995) Social Commitment: from individual intentions to groups and organizations. In *ICMAS'95 First International Conference on Multi-Agent Systems*, AAAI-MIT Press, 41–49.

Castelfranchi, C., Giardini, F., Lorini, E., Tummolini, L. (2003) The prescriptive destiny of predictive attitudes: from expectations to norms via conventions, in R. Alterman, D. Kirsh (eds.) *Proceedings of the 25th Annual Meeting of the Cognitive Science Society,* Boston, MA.

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

Conte, R. and Paolucci, M. (2002) *Reputation in Artificial Societies. Social Beliefs for Social Order*. Boston: Kluwer Academic Publishers.

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.

Festinger, L. (1957) *A Theory of Cognitive Dissonance.*, Stanford, CA: Stanford University Press.

Gintis, H. Strong reciprocity and human sociality, *Journal of Theoretical Biology*, 206: 169–179, 2000.

Yu, B. and Singh, M. P. (2003) Searching social networks. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*. Pp. 65–72. ACM Press.

Garfinkel, H. (1963) A conception of, and experiments with, 'trust' as a condition of stable concerned actions. In O. J. Harvey (ed.) *Motivation and Social Interaction*. Ronald Press, NY, Ch. 7, pp. 187–238.

Giddens, A. (1984) *The Constitution of Society: Outline of the Theory of Structuration*. Cambridge, Cambridge University Press.

Goffman, E. (1959) *The Presentation of Self in Everyday Life*. Anchor Books.

Pendenza, M. *Introduzione* a Harold Garfinkel 'La fiducia', (Italian trnslation of H. Garfinkel, 1963), Armando, Roma, 2000.

Searle, J.R. (1995) *The Construction of Social Reality*, London: Allen Lane.

Tummolini, L. and Castelfranchi, C. The cognitive and behavioral mediation of institutions: Towards an account of institutional actions. *Cognitive Systems Research*, 7 (2–3), 2006.

Thomas de Aquino, Summa Theologiae (http://www.corpusthomisticum.org/sth1084.html)