Trust revisited

Andrew J I Jones

King’s College London, Dept of Computer Science

andrewji.jones@kcl.ac.uk

Some current definitions

- Examples taken from Castelfranchi & Tan, eds., *Trust and Deception in Virtual Societies,* (Kluwer 2001).
- In general, 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. (Bacharach & Gambetta)

Some current definitions (contd)

- In the context of a society, trust can be viewed as a mechanism for reducing complexity and a means of coping with the freedom of others - it is an aspect of all social relationships and implies some form of expectation about the future. Barber breaks trust down into three components: *(1)* an expectation of the fulfilment of the natural social order, *(2)* an expectation of competent role performance on the part of the trustee and *(3)* an expectation that a trustee will fulfil all fiduciary obligations. (Rea)

Some current definitions (contd)

- Like many other definitions of trust, Zaltman and Moorman define trust through prediction that is value free: "an interpersonal or interorganizational state that reflects the extent to which the parties can predict one another's behavior; can depend on one another when it counts; and have faith that the other will continue to act in a responsive manner despite an uncertain future." This definition does not address what we often assume to be characteristic of trust - that those expectations are largely about outcomes that are in common with our own interests. (Elofson)

Some current definitions (contd)

- ....Giffin includes trust's implicit goal-directed characteristic in providing an alternative definition: "reliance upon the characteristics of an object, or the occurrence of an event, or the behavior of a person in order to achieve a desired but uncertain objective in a risky situation." Further, she cited the following elements as essential to describing a trusting person:

➢ A person is relying on something.

➢ This something relied upon may be an object, an event, or a person.

➢ Something is risked by the trusting person.

➢ The trusting person hopes to achieve some goal by taking this risk.

➢ The desired goal is not perceived as certain.

➢ The trusting person has some degree of confidence in the object of his trust. (Elofson)

Some current definitions (contd)

- Using these definitions, a composite definition of trust was then suggested: trust is the outcome of observations leading to the belief that the actions of another may be relied upon, without explicit guarantee, to achieve a goal in a risky situation. (Elofson)

An identifiable core ?

- Despite the diversity in recent attempts to define *trust*, there is perhaps a core common to most of them, which can be located in the notion of *expectation*. (See *On the concept of trust*, DECISION SUPPORT SYSTEMS, 2002.)
- Try to make this idea more precise, by analysing four examples of situations in which it would ordinarily be said that one agent trusts another.

Risk, Dependence, Intended Goal

- The examples will also be used to comment on the relationship between *risk* and *trust*.
- Furthermore - in part in contrast to the approach of Castelfranchi and Falcone – it will be suggested that the notion of *intended goal*, is not a **necessary** feature of *trust*.

Four scenarios

- In each of the following situations it is true to say that *x* trusts *y*:

S1- (*the regularity scenario*)

- *X* believes that there exists a regularity in *y*'s behaviour, so that under particular kinds of circumstances *y* exhibits a particular kind of behaviour (he does *Z*). In addition, *x* believes that this regularity will persist.

S2- (*the obligation scenario*)

- *X* believes that there is a rule requiring *y* to do *Z*, and that *y*'s behaviour will in fact comply with this rule. For instance, *x* believes that *y* is under an obligation to repay a debt, and that *y* will indeed make the repayment.

Four scenarios (contd)

S3 – (*the role scenario*)

- *X* believes that *y* occupies some particular role, and that *y* will perform the acts associated with that role in a competent manner. This is what is meant when it is said, for instance, that *x* trusts his doctor, or *x* trusts his car mechanic.

S4 – (*the informing scenario*)

- *X* believes that *y* is transmitting some information to him, and that the content of *y*’s message, or signal, is reliable. For instance, *y* says to *x* “Norwegians eat rotten fish”, and *x* believes what he says.

An analysis of the examples

- In each of these scenarios, the core of *x*'s trusting attitude lies in two beliefs, which will be called *the rule-belief* and *the conformity-belief*, respectively.

An analysis of the examples (contd)

S1- (*the regularity scenario*)

- In S1, the rule-belief is *x*'s belief that there exists a regularity in *y*'s behaviour.
- *x*'s rule-belief: $B_x(A \approx> E_y Z)$
- x’s conformity belief is that the regularity will persist, so that he may continue to draw the default conclusion that $E_y Z$, when he believes that A occurs.

An analysis of the examples (contd)

- In distinguishing cases of type S1 from cases of type S2, it should be noted that there is in the former, as here understood, no assumption of an agreement between *x* and *y*, or of the existence of an obligation, according to which *y* is *required* to do *Z*. This feature of cases of type S1 might be described by saying that *x*'s expectation vis-à-vis *y* is a purely *factual* - rather than *normative* - expectation. If *y* does not do *Z*, and thus fails to act in accordance with *x*'s expectation, *x* will see this as an *exception* to the believed regularity in *y*'s behaviour and not as an act of *violation* of some obligation.

An analysis of the examples (contd)

S2- (*the obligation scenario*)

- In S2, the rule-belief component of *x*'s trusting attitude is *x*'s belief that *y* is under an obligation to do *Z*. And the conformity-belief component is *x*'s belief that *y* 's behaviour will be of a kind which fulfils this obligation. Here, *x* may be said to have a normative expectation vis-à-vis *y* in the sense that *x* believes that there is a requirement that *y* is to do *Z*. This expectation - *x*'s rule-belief - is in itself compatible with a belief, or suspicion, on *x*'s part that *y* will violate his obligation; however, *x*'s conformity-belief is that what in fact will happen is that *y* will meet his obligation, i.e., that *y* will do what he is supposed to do. In cases of type S2, then, trust amounts to belief in *de facto* conformity to normative requirement.

An analysis of the examples (contd)

S2- (*the obligation scenario contd*)

- Rule-belief: B_xOE_yZ
- Conformity-belief: $B_x(OE_yZ \text{ -> } E_yZ)$

An analysis of the examples (contd)

S3 – (*the role scenario*)

- Scenarios of type S3 are intended to cover such uses of *trust* as are exemplified by "*x* trusts his doctor", "*x* trusts his car-mechanic", and so on. The assumption is that what is said to be trusted in these instances is behaviour associated with some particular role(s): *x* trusts his doctor/car-mechanic to perform competently the roles associated with being a doctor/being a car mechanic.

An analysis of the examples (contd)

S3 – (*the role scenario contd*)

- The rule-belief/conformity-belief model again applies: a central feature of any given role is that it has associated with it a set of normative standards. It is required of a doctor, for instance, that he exercise particular skills in ways which meet certain standards of competence. The rule-belief component of *x*'s trust in his doctor *y*, is *x*'s belief that there are standards that the actions of an agent occupying the role of doctor are required to meet. The conformity-belief is *x*'s belief that *y*'s actions will satisfy these standards.
- Thus, on this approach, scenarios of type S3 turn out to be particular instances of scenarios of type S2.

An analysis of the examples (contd)

- S4 – (*the informing scenario*)

- Does a scenario of type S4 represent a quite different type of trust from that identified for S2 and S3 ? The answer depends on how the communicative act-type of *saying, stating or asserting that such-and-such* (the indicative signalling act-type) is to be characterised.

- The approach adopted here is that an indicative signalling-system is constituted by rules, or conventions, which grant that the performance, in particular circumstances, of instances of a given class of act-types *count as* assertions; and these rules also specify what the assertions mean.

An analysis of the examples (contd)

S4 – (*the informing scenario, contd.)*

- For example, the utterance of a token of the sentence "The ship is carrying explosives" will count, in an ordinary communication situation, as an assertion that the ship is carrying explosives. The raising, on board the ship, of a specific sequence of flags, will also count as an assertion that the ship is carrying explosives.
- In the first case signals take the form of linguistic utterances, and in the second they take the form of acts of showing flags. But for both signalling systems there are rules determining that particular acts count as assertions with particular meanings.

An analysis of the examples (contd)

- S4 – (*the informing scenario, contd.)*
- According to Searle, if the performance by agent *y* of a given communicative act counts as an assertion of the truth of *p*, then *y*'s performance *counts as an undertaking to the effect that p is true* . What lies behind that claim, surely, is that, when *y* asserts that *p*, what he says *ought* to be true, in some sense or other of 'ought'. The problem is to specify what sense of 'ought' this is.
- The view accepted here is that the relevant sense of 'ought‘ is like that used in "The meat ought to be ready by now, since it has been in the oven for 90 minutes." The system (oven with meat in it) is functioning less than optimally if it is not ready - things are then not as they ought to be, something has gone wrong.

An analysis of the examples (contd)

S4 – (*the informing scenario, contd.)*

- The rule-belief/conformity-belief model applies for the S4 cases too, but the rules the truster (*x*) believes to hold are those which specify what should optimally be the case when an indicative signal is transmitted. When *x* trusts the reliability of *y*'s assertion - when he trusts what *y* says - he also believes (conformity-belief) that *y*'s act satisfies the optimality condition embodied in the governing signalling rule.
- Rule-belief: $B_x(E_yC \ \ {=}{>}_s\ I^*{}_sA)$
- Conformity-belief: $B_x(E_yC \text{ -> } A)$

An analysis of the examples (contd)

- It might be maintained that the rules which are the objects of rule-beliefs in scenarios of type S2 and S3 express *norms*, whereas the rules which are the objects of rule-beliefs in scenarios of type S4 describe signalling *conventions*. Classifying matters in this way would be perfectly acceptable. It is clear that my use of the term 'rule-belief' trades on the ambiguity of the term 'rule', which may pertain (S1) to a regularity, or (S2 and S3) to a norm, or (S4) to a convention.
- The main point is that the attitude of trust may in each case be understood as a belief in conformity to a "believed-in" rule: the fact that the modal status of the rule may be different in different types of scenarios does not undermine the claim that there is an identifiable common core to the meaning of 'trust'.

Trust and *risk*

- Some would maintain that the term *trust* is appropriately used to describe the four scenarios only if we *also* suppose that the truster is exposed to *risk* if the trustee fails to conform.
- In other words, for genuine trust, the truster has an *interest in trustee compliance.*

Trust and *Risk* (contd)

- If we accept a claim of this sort, it may be accommodated within the formal framework here proposed by adding an additional type of normative modality – an *evaluative* normative modality – to represent what an agent takes to be in his/her interests (or, perhaps, what he/she prefers).

Evaluative and directive normative modalities

- Represent ‘*x* deems *A* to be in his/her interests’ by $I_x A$
- Note the distinction between this evaluative normative modality and the *directive* modality (above represented by the O-modality) used to represent what is *obligatory for* or *required of* an agent.

Evaluative and directive normative modalities (contd)

- A distinction of roughly this sort goes back at least to Kanger, and was exploited by Pörn in his Action Theory and Social Science (Reidel, 1977) to articulate the difference between *wants* and *intentions*, where the former are expressed in terms of the evaluative modality, whilst the latter are interpreted as the directives an agent adopts to steer his/her own actions.

Further comments

- There is a tendency, in some of the recent literature on trust, for *intention* to be defined in terms of an evaluative modality (*preference*).
- See Herzig et al., *Prolegomena for a logic of trust and reputation*, and Hűbner et al., *From cognitive trust theories to computational trust.*

Further comments (contd)

- This in part comes about because of the authors’ stated aim of formally modelling the Castelfranchi & Falcone intention-based analysis of trust.
- In my view, as here outlined, the notion of intention is largely irrelevant to the definition of trust.