Socio-Cognitive Model of Trust

Rino Falcone
Institute for Cognitive Sciences and Technology, National Research Council of Italy, Italy

Cristiano Castelfranchi
Institute for Cognitive Sciences and Technology, National Research Council of Italy, Italy

INTRODUCTION

Humans have learned to cooperate in many ways and in many environments, on different tasks, and for achieving different and several goals. Collaboration and cooperation in their more general sense (and, in particular, negotiation, exchange, help, delegation, adoption, and so on) are important characteristics - or better, the most foundational aspects - of human societies (Tuomela, 1995).

In the evolution of cooperative models, a fundamental role has been played by diverse constructs of various kinds (purely interactional, technical-legal, organizational, socio-cognitive, etc.), opportunely introduced (or spontaneously emerged) to support decision making in collaborative situations.

The new scenarios we are destined to meet in the third millennium transfigure the old frame of reference, in that we have to consider new channels and infrastructures (i.e., the Internet), new artificial entities for cooperating with artificial or software agents, and new modalities of interaction (suggested/imposed by both the new channels and the new entities). In fact, it is changing the identification of the potential partners, the perception of the other agents, the space-temporal context in which interaction happen, the nature of the interaction traces, the kind and role of the authorities and guarantees, etc.

For coping with these scenarios, it will be necessary to update the traditional supporting decision-making constructs. This effort will be necessary especially to develop the new cyber-societies in such a way as not to miss some of the important cooperative characteristics that are so relevant in human societies.

BACKGROUND

Trust (Ganbetta, 1990; Luhmann, 1990; Dasgupta, 1990), in the general frame described above, might be considered as a socio-cognitive construct of main importance. In particular, trust building is always more recognized as a key factor for using and developing the new interactional paradigm.

Trust should not be made indistinct with security. The latter can be useful to protect – in the electronic domain - from the intrusiveness of an unknown agent, to guarantee an agent in the identification of its partner, to identify the sender of a message (for example, by verifying the origin of a received message; by verifying that a received message has not been modified in transit; by preventing that an agent who sent a message might be able to deny later that it sent the message [He, Sycara & Su, 2001]). With sophisticated cryptographic techniques, it is possible to give some solution to these security problems.

However, more complex is the issue of trust, that must give us tools for acting in a world that is in principle insecure (that cannot be considered 100% secure), where we have to make the decision to rely on someone in risky situations. (Consider the variety of cases in which it is necessary or useful to interact with agents whose identity, history or relationships are unknown, and/or it is only possible to make uncertain predictions on their future behaviors.)

Trust should not be made indistinct with reputation, too. In fact, communicated reputation (Conte & Paolucci, 2002) is one of the possible sources on which the trustier bases its decision to trust or not.

The more actual and important example of the usefulness of trust building is electronic commerce, but we must also consider other important domains of Multi Agent Systems and Agent Theory such as Agent Modeling, Human-Computer Interaction, Computer Supported Cooperative Work, Mixed Initiative and Adjustable Autonomy, Pervasive and Ubiquitous Computing. In fact, today many computer applications are open distributed systems (with many autonomous components that are spread throughout a network and interacting with each other). Given the impossibility to rule this kind of system by a centralized control regime (Marsh, 1994), it becomes essential to introduce local tools in order to choose the right partnership and at the same time reduce the uncertainty (deriving from the nature of an open distributed system) associated with that choice.

TRUST IN THE NEW TECHNOLOGICAL SCENARIOS

In fact, various different kinds of trust should be modeled, designed, and implemented:

- Trust in the environment and in the infrastructure (the socio-technical system)
- Trust in personal agents and in mediating agents
- Trust in potential partners
- Trust in sources
- Trust in warrantors and authorities.

Part of these different kinds of trust have a complementary relation with each other, that is, the final trust in a given system/process can be the result of various trust attributions to the different components. An exemplary case is one's trust in an agent that must achieve a task (and more specifically in its capabilities for realizing that task) as different from one's trust in the environment (hostile versus friendly) where that agent operates, or again as different from one's trust in a possible third party (arbitrator, mediator, normative systems, conventions, etc.) able to influence/constrain the trustee and representing a guaranty for the trustier (Castelfranchi & Falcone, 1998; Falcone & Castelfranchi, 2001).

Therefore, the "sufficient" trust value of one single component cannot be established before evaluating the value of the other components. In this regard, it is very interesting to characterize the relationships between trust and (partial) control (Castelfranchi & Falcone, 2000).

It is important to underline how trust is in general oriented towards not directly observable properties. It is, in fact, based on the ability to predict these properties and to rely or not to rely on them. Thus, it is quite complex to assess the real trustworthiness of an agent/system/process, not only because - as we have seen - there are many different components that contribute to this trustworthiness, but also because the latter is not directly observable (see [Bacharach & Gambetta, 2001] about signs of trust). The important thing is the perceived trustworthiness that is, in its turn, the result of different modalities of the trustier's reasoning about direct experience; categorization; inference, and communicated reputation.

SOCIO-COGNITIVE MODEL OF TRUST

The Socio-Cognitive model of trust is based on a portrait of the mental state of trust in cognitive terms (beliefs, goals). This is not a complete account of the psychological dimensions of trust. It represents the most explicit (reason-based) and conscious form. The model does not account for the more implicit forms of trust (for example, trust by default, not based upon explicit evaluations, beliefs, derived from previous experience or other sources) or for the affective dimensions of trust, based not on explicit evaluations but on emotional responses and an intuitive, unconscious appraisal (Thagard, 1998).

The word *trust* means different things, but they are systematically related with each other. In particular, three crucial concepts have been recognized and distinguished not only in natural language but also in the scientific literature. Trust is at the same time:

- A mere *mental attitude* (prediction and evaluation) toward another agent, a simple *disposition*;
- A *decision* to rely upon the other, i.e., an *intention* to delegate and to trust, which makes the trustier "vulnerable"(Mayer, Davis, & Schoorman, 1995);
- A *behavior*, i.e., the intentional *act* of trusting, and the consequent *relation* between the trustier and the trustee.

In each of the above concepts, different sets of cognitive ingredients are involved in the trustier's mind. The model is based on the BDI (Belief-Desire-Intention) approach for modeling mind that is inspired by Bratman's philosophical model (Bratman, 1987). First of all, in the trust model only an agent endowed with both goals and beliefs can "trust" another agent. Let us consider the trust of an agent *X* towards another agent *Y* about the (*Y*'s) behavior/action α relevant for the result (goal) *g* when:

- *X is the (relying) agent, who feels trust; it is a cognitive agent endowed with* internal explicit goals and beliefs (the *trustier*)
- *Y* is the agent or entity that is trusted (the *trustee*)
- *X* trusts *Y* about g/α and for g/α.

In the model *Y* is not necessarily a cognitive agent (for instance, an agent can - or cannot - trust a chair as far as to sustain his weight when he is seated on it). On the contrary, *X* must always be a cognitive agent: so, in the case of artificial agents we should be able to simulate these internal explicit goals and beliefs.

For all the three notions of trust defined above (*trust disposition*, *decision to trust*, and *trusting behavior*), we claim that someone trusts some other one only relatively to some goal (here the goal is intended as the general, basic teleonomic notion, any motivational representation in the agent: desires, motives, will, needs, objectives, duties, utopias, are kinds of goals). An unconcerned agent does not really "trust": he just has opinions and forecasts. Second, trust itself *consists of* beliefs.