Know Techn Pol (2010) 23:347–366
DOI 10.1007/s12130-010-9118-4

SPECIAL ISSUE

What Is the Model of Trust for Multi-agent Systems? Whether or Not E-Trust Applies to Autonomous Agents

Massimo Durante

Received: 22 March 2010 / Accepted: 18 July 2010 / Published online: 18 September 2010

Abstract A socio-cognitive approach to trust can help us envisage a notion of networked trust for multi-agent systems (MAS) based on different interacting agents. In this framework, the issue is to evaluate whether or not a socio-cognitive analysis of trust can apply to the interactions between human and autonomous agents. Two main arguments support two alternative hypothesis; one suggests that only reliance applies to artificial agents, because predictability of agents' digital interaction is viewed as an absolute value and human relation is judged to be a necessary requirement for trust. The other suggests that trust may apply to autonomous agents because predictability of agents' interaction is viewed only as a relative value since the digital normativity that grows out of the communication process between interacting agents in MAS has always deal with some unpredictable outcomes (*reduction of uncertainty*). Furthermore, human touch is not judged to be a necessary requirement for trust. In this perspective, a diverse notion of trust is elaborated, as trust is no longer conceived only as a relation between interacting agents but, rather, as a relation between cognitive states of control and lack of control (*double bind*).

Keywords Trust · Multi-agent system · Uncertainty · Autonomous agent · Cognitive states · Normativity

1 Introduction

This paper consists of four sections except for the introduction (Section 1). In Section 2, the issue of building trust with regards to the nature of digital environments and the status of interacting agents is taken into consideration: a multi-agent system is to account for a comprehension of trusting interactions that are field-dependent and vary according to the heterogeneity of interacting agents. In this

M. Durante (✉)
Law School, University of Torino, Turin, Italy
e-mail: massimo.durante@unito.it

section, we lay down three principles, which govern the process of building trust. Such process requires rethinking the notion itself of trust in digital environments. In Section 3, we sketch out a socio-cognitive notion of trust (along the guidelines traced by Castelfranchi and Falcone 2008), which deals with the problem of elaborating forms of online cooperation among interacting parties. To this end, we analyse the key elements that form such notion of trust and we pay attention to the features that characterise both the trustor and the trustee. This analysis allows us to revise some common characteristics that are usually associated with trust; in Section 4, we propose to look at the notion of trust from the standpoint of the principle of uncertainty, which is of a growing importance in the epistemological arena (Halpern 2005). This approach to trust turns out to be less anthropocentrically biased than a socio-cognitive analysis would suggest. In this regard, we take advantage of Luciano Floridi's (2001, 2004) ethics of information. A less anthropocentric approach enables us, in Section 5, to apply our study of trust to the relation between human and autonomous agents. To this aim, we investigate the main features of AAs from a cognitive viewpoint and discuss the chief arguments, which contrast the feasibility and advisability of trusting AAs. Our analysis is far from being exhaustive: however, it is possible, on this basis, to elaborate further cognitive models, which attempt to work out some of the conceptual difficulties that affect trusting AAs. In this perspective, we designate four main tenets that appear crucial for a theoretical and practical development of AAs as regards to networked trust: (1) the traceability of a decision-making process; (2) the informational feedback; (3) the process of assigning variables to values; (4) the degrees of control and lack of control. From the start, we have to clarify what we mean by *control*. We do not refer to any authority, express influence or power to direct or determine behaviours. As in control theory, we aim to realize how unstable processes can be stabilized, but our investigation is limited to the domain of trust, and it refers only to the cognitive and epistemic situation where we hold relevant and reliable, but incomplete information about a given reality. Thanks to that information (and notably to information feedbacks) that reality can be subjected to our evaluation and therefore, in this narrower sense, to our control. Being in command of a situation means, in these terms, being able to provide our decisions and behaviours with a cognitive and rational basis made of (as much as possible relevant and reliable) information.

2 Field Dependency and Heterogeneity of Agents in MAS

Since cyberspace supports a variety of agents, from human to artificial agents, when interacting in the multi-agent system, we rely more and more on the human/machine interface for communication, information, transactions, business, without being always able to ascertain whether we interact with human agents or with artificial (autonomous) agents. Furthermore, all the major aspects of multi-agent systems require not only reliance on the machine interface protocols but, first and foremost, trust in someone else's behaviour for the success of operations, negotiations and relations based on computer-mediated interaction or coordination between individuals or groups. From this perspective, I suggest that our normative understanding of possibly trustworthy interactions is *field-dependent*, i.e. it depends on the context it occurs in, and varies according to the nature of the agents we interact with.

Therefore, the question of trust-building in web technologies-based environments depends on what is the type of normativity in the context of interaction and on what the nature is of the interacting agents: human/human (H>H); human/agents (H>A); agents/human (A>H); agents/agents (A>A).

With regard to the normativity of the context, I suggest that building trust is not only a matter of assuring technological or legal security by means of rules, constraints, protocols, architectures, and guaranties. Technological or legal security is particularly necessary in the context of online commercial transactions, privacy issues, and legal contracts, but it does not suffice to assure that trustworthy interactions are displayed in the context of social cooperation. In fact, this is to be based on the agent's behaviour, on its willingness to cooperate with us that is never fully predictable nor can it be automatically subject to control or implementation. On the contrary, it is mainly concerned with mental and social dispositions towards other agents that are to be envisaged within the framework of a model of networked trust.

To be more precise, I can state three normative principles that govern the process of building trust, that is, how we try to secure that a particular goal is obtained by means of someone else's behaviour:

- *Principle of compensation*=legal security is required as long as agents' interaction can be based on a rule of compensation: this means that the agents' desired goal can be efficiently substituted by a second-best goal. In this regard, law is a system of expectations and of rules compensations, which apply when expectations are frustrated. In other words, one interacting party can provide the other with compensation in case of defection. From a cognitive standpoint, the legal system is based on expectations as well as a trust system but it assures compensation over trust, firstly. For this reason, it is actually said to be based on parties' mutual distrust (Chiodi 2000; Scillitani 2007; Resta 2009), even though distrust is ultimately based on parties' trust in the effectiveness of a third authority, which is called upon to apply the rules of compensation (Durante 2008);
- *Principle of prevention*= technological security is required as long as interacting parties believe that agent's expected behaviour does not suffice by itself to secure the desired goal, nor are they satisfied with a second-best goal assured by means of compensation. One party intends to prevent the other from not fulfilling the expected task: this system assures prevention over trust. It is actually based on the parties' common distrust, since all parties share distrust in the effectiveness of a third authority called upon to apply the rules of compensation. From this perspective, trust is displaced from parties or authorities and is placed on technological devices (at times abusively but significantly called *trusted systems*). However important, prevention should not always prevail over trust: "My own preference would be for a progressive social vision of cyberspace that preserves the degrees of freedom that trust needs. At the same time, we ought to develop technologies of security that might make possible pockets of high security for the kinds of transactions that call for it, without making that the *dominant norm* throughout" (Nissenbaum 2004, p. 179, [emphasis added]);
- *Principle of cooperation*=social trust is required as long as agents' interaction cannot be based either on rules of compensation nor on trusted systems of prevention. One interacting party cannot (or prefers not to) obtain a desired goal if

not relying on the other party's will and ability to fulfil the delegated task, that is, taking the risk to entrust someone else. This system assures trust over security. It is based upon a mutual relation of common trust: "The key behaviour of the agents to enable them to form cooperative group is that they shift their probability of cooperation or defection based on the *expected behaviour* of the majority of its neighbours, i.e. if the majority of neighbours play defect then each agent will increase the probability that it defects, and the same for cooperation" (Ghanea-Hercock 2007). Furthermore, expected behaviours are reinforced by sharing common concerns for a goal: "Trust-based online cooperation proves to be reasonable whereas it is teleologically aimed to a specific goal of common concern grown out of a communication process drawing the framework within which it is reasonable to expect a determined behaviour from another agent" (Durante 2008).

One example can explain how these principles can apply according to a different level of abstraction (LoA, see Floridi 2008). Let us imagine the relation between a producer and a musical band. Firstly, the producer is interested in the fact that the musical band will perform, during the year, a certain number of concerts. She will make the band sign a contract, in order to provide herself with compensation, in case of defection (*principle of compensation*). Secondly, the producer is interested in the fact that the band will perform a concert, even if the band is not in a fit condition for playing. She will make the band perform the concert by means of playback, in order to prevent them from not fulfilling their expected task (*principle of prevention*). Thirdly, both the producer and the band are interested in showing that the band is a great live performer: all of them need to trust each member of the musical band that they will be able to perform the common task (*principle of cooperation*).

To review what has been outlined so far, I can state that different levels of abstraction or field dependency might require hence a triple shift of paradigm in the study of networked trust insofar as the functioning of MAS demands cooperation:

1. From technologically gained security to perceived security (security as it is perceived by agents disposing of incomplete information);
2. From control trust (trust based on control tools and mechanisms for assessing trustworthiness) to party trust (trust based on the dynamic interaction between a party, the trustor, and a counter party, the trustee);
3. From a model of probabilistic trust (based on rigid methods of statistical inferences) to a model of cognitive social trust (based on beliefs, expectations and concerns).

3 A Socio-Cognitive Model of Trust

This triple shift of paradigm is accomplished by the cognitive model of social trust elaborated by Castelfranchi and Falcone (2007, 2008), which I have already expounded in a previous essay (Durante 2008). Let me restate the main conditions and elements of the model.

Building cooperation upon trust requires taking into account the following conditions:

1. *Role of expectations*: trustor intention to trust is governed by the aim of *stabilizing expectations* (the normative status of trust), so that it becomes crucial to the role of beliefs, mental representations and expectations of the trustor, which result from a degree of trustworthiness;
2. *Lack of certainty*: the amount of information that trust rests on is neither fully complete nor fully incomplete: we can neither speak of an absolute absence of information (some elements are required, in order to evaluate trustworthiness) nor of certain information (which would exclude the risk inherent to the act of trusting). Trust is governed by a *principle of uncertainty* (the informational status of trust), which makes the decision to trust possible.
3. *Inherent risk*: trust is meant to exist insofar there is a risk (Luhmann 1979) or at least something beyond control that is something that the trustor cannot entirely be in command of (either from a cognitive or practical point of view). What *beyond control* is is not necessarily at variance with the idea of rationality (the rational status of trust). On the contrary, rationality has always to deal with what remains *to some extent* beyond control, the problem being how to cognitively represent this extent (much on this in the Section 4).

Trust consists of three basic elements (Castelfranchi and Falcone 2008):

- A *mental attitude*, a pre-disposition towards another agent: this attitude is a belief (the strongest cognitive element), constituted by the evaluation of the agent's trustworthiness and a prediction regarding the agent's willingness and ability to produce some effect;
- A *decision to rely* upon the other: according to us, this decision always brings forth the intention to delegate the production of a desired goal, which exposes the trustor to a risk and makes them vulnerable;
- A *behaviour*: the effective act of entrusting another agent that entails a practical, informational relation between the parties.

Trust consists of a stratified dimension made up of: (Section 3.1) a trust attitude (TA); (Section 3.2) a decision to trust (DtT); and (Section 3.3) an act of trusting (AoT). Socio-cognitive trust is not, however, only a mental attitude, or a pure internal belief. The context (i.e. the informational environment) must also be taken into account, where the trustor enters into relation with the trustee, and which affects the evaluation, prediction and decision to trust, and where the trustee is meant to produce the desired goal, which affects the possibility of success.

3.1 The TA: The Role of Beliefs and Context

In order to trust in someone, the trustor should have positive expectations both about (a) the probability of the trustee performing the task well (internal attribution) and about (b) the external conditions that might influence the trustee's actions (external attribution).

1. Internal attribution

"Trust is not simply a prediction (although it implies a prediction). In particular, it is not a prediction of a given behaviour based on some observed frequency, on some estimated probability, or on some regularity and norm. It requires the grounding of such a prediction (and hope) on an internal attribution" (Castelfranchi and Falcone 2007). The internal attribution of trust depends on the evaluation of the trustee's qualities and defects that define trustworthiness and ground the basic beliefs of trust (in Section 5, I will analyse to what extent autonomous agents can endorse such qualities and defects, which are not necessarily properties of human beings but, more generally, of trusting cognitive systems):

- *Competence*: the set of qualities that makes the trustee able to perform the task. They are internal qualities of the trustee such as skills, know how, expertise, knowledge, self-confidence;
- *Willingness*: the trustee's real intention (will) and readiness (organization) to perform the task;
- *Persistence*: the trustee's attitude of steadiness in their intention to perform the task;
- *Dependence*: the trustor's belief that it is either necessary (strong dependence) or preferable (weak dependence) to rely on the trustee in order to obtain a goal;
- *Fulfilment*: the trustor's belief that the goal will be achieved thanks to the trustee, since the latter clearly understands what the goal consists in;
- *Motivation*: the set of reasons that can persuade the trustee to adopt the goal. They are crucial factors in trusting since they constitute specific motives to perform the task such as concern for the goal, friendship, altruism, reciprocity, and cooperation.

2. External attribution

External attribution of trust depends on the trustor's evaluation of the appropriate environmental conditions that might influence the trustee's actions and, therefore, their performance and success. External trust is related to two types of condition (in Section 5, I will analyse to what extent autonomous agents can deal with these conditions):

   – Positive conditions: concerning the presence of opportunities and resources;
   – Negative conditions: concerning the absence of interferences and adversities (uncertainty; absence of controls or resources; situation of risk or danger; lack of security).

The external attribution of trust plays a systemic role in networked cooperation since it allows the trustor to cope with the complexity of environment, which can be faced along these guidelines:

- Favourable environmental conditions:

   – The more favourable the perceived environmental conditions are, the smaller the need for trust toward the trustee, and vice versa;
   – The more perceivable and apt to manipulation are the environmental conditions, the more AA are likely to play a crucial role in the multi-agent system.

- Adverse environmental conditions:
  - The more adverse the environmental conditions, the more we are expected to make use of a sharpened and tighter internal attribution of party trust;
  - The higher the perceived internal attribution of trust, the more possible it is to deal with *invisible* (not entirely perceived: Moor 1985) adverse environmental conditions;

Internal and external trust both affect the trust attitude, but they do not affect the trust decision. It is precisely that decision which allows the trustor to reduce the complexity of the environment in relation to the system of networked cooperation.

3.2 The DtT: Scalable Evaluation and Clear-Cut Decision

As we have said above, the trust attitude consists of a set of mental beliefs by means of which the trustee is evaluated. This cognitive evaluation contains several degrees: the level of trust is dependent on various conditions and factors; such as the trustee's reliability, the evaluation of external conditions, the probability of the trustee's action and success etc. These conditions and factors may be perceived, rationally evaluated and quantified: they represent what is known by experience or at least perceived by the trustor in the situation of uncertainty and risk he is exposed to. Trust attitude consists of a scalable evaluation: degrees of trust are related to rational beliefs and this allows the trustor to evaluate the risk and the advisability of trusting: for instance, 0 may represent the null degree of trust, while 1 may represent the full degree of trust. A calculation can be made, although not only a merely probabilistic one, so that we can say that trust may be attributed in a stronger or weaker manner.

This should not, however, cause us to lose sight of the fact that the decision to trust is a *clear-cut decision*: when the trustor is ultimately faced with the possibility of trusting, he decides either to trust or not to trust. This is the crucial aspect of trust: the decision to trust stems from a rational cognitive calculation but it is not a calculation in itself. The decision to trust affects the trustors' status in their integrity, which is not a scaling or relative concept. The clear-cut dimension of the decision to trust is an inherent character of a trust relationship (whatever be the agent).

In the trust attitude, the otherness of the agent (i.e. whether or not the task will be performed) is reduced to a cognitive evaluation: it is subject to reduction, namely it becomes something else, that is to say the subject of an evaluation and therefore of a calculation (for instance, a number between 0 and 1). In the trust decision, the otherness of the agent is not reduced to something else: the trustor has to enter into a relationship with the otherness of the agent, namely with what can never entirely be evaluated and therefore calculated. When entering into a relationship with the other agent through the decision to trust, the trustor establishes communication with the trustee. This communication is affected by the decision to trust, which gives rise to both information (about the trustor's reliance and delegation) and a norm (the trustee's actions will be judged according to the decision to trust). Trust is not only the means by which trustors can take into account the otherness of the agent, but it is above all the means by

which they can enter into a form of communication, through which they exchange information.

This exchange of information (to be tracked) governs the relationship between agents, whether human or artificial and, within the artificial dimension, whether autonomous or not, according to the criteria established by Luciano Floridi (2004, 358), at a particular level of abstraction (Floridi 2008), in relation to the agent's attitudes of interactivity, autonomy and adaptability, towards information and rules.

The right LoA is probably one that includes the following three criteria: (1) interactivity, (2) autonomy and (3) adaptability.

1. Interactivity means that the agent and its environment (can) act upon each other. Typical examples include input or output of a value, or simultaneous engagement of an action by both agent and patient—for example gravitational force between bodies.
2. Autonomy means that the agent is able to change state without necessitating a direct response to an interaction: it can perform internal transitions to change its state. So an agent must have at least two states. This property imbues an agent with a certain degree of complexity and independence from its environment.
3. Adaptability means that the agent's interactions (can) change the transition rules by which it changes state. This property ensures that an agent might be viewed, at the given LoA, as learning its own mode of operation in a way which depends critically on its experience.

3.3 The AoT: Teleological Dimension of Networked Trust

It is possible to think of trust as unconditional or even blind, but according to a socio-cognitive model of trust, it is rationally possible to trust only in relation to a goal. As seen in Section 2, the normativity of trust (i.e. principles of compensation, prevention and cooperation) is set in relation to a goal to be obtained. Trust is not general or indiscriminate credit or allegiance given to someone. It must always be measured with reference to a specific objective (conditional trust): this makes crucial the specification of goals by the trustor. This is also relevant in both an epistemic and a pragmatic sense from the trustee's point of view, since the trustee must understand what the goal really is and show (HA) or have (AA) concern for its achievement. The trustor may trust the agent if she thinks or perceives that the agent has grasped what the delegated task and the desired goal is. The trustee's concern for the goal is not only part of the agent's motivation but also ensures that the agent knows what to do. These further requirements (mutual understanding, concern and goal) narrow the domain of trusting networked cooperation, which is more likely to be effectively achieved when the agents share:

- *A common knowledge and understanding of the epistemic structure of the task to be performed*: online cooperation is likely to work where common interests already exist or at least where there is a high level of sharing of knowledge;
- *A common goal to be achieved*: the more detail in which the goal is specified, the more the agent can anticipate and adapt itself to obtain it. In addition to this, the more the goal is achieved by means of modular contributions, the more it is

necessary to rely on a great number of contributions: in this case, however, it is less harmful if some do not fulfil their task;
- *A common concern for the goal to be achieved*: online cooperation in MAS often requires goal-directed interactivity. This cooperation can be based on incremental contributions: in this case the decision to trust is not to be based merely on individual willingness to perform the immediate task, but also on shared concern for the final goal to be achieved.

Networked cooperation based on trust inevitably has a *teleological* dimension: it is concerned with a goal to be achieved, which is the underlying reason for delegation and action. Trust is primarily based on predictions, but not entirely, since predictions will be adjusted according to the actual behaviour that agents display in relation to a goal. The trustee's behaviour gives the trustor new information that may lead to changes in their predictions. This information is only made possible by two conditions, namely by two forms of profiling/specification:

- The previous specification of goals: *profiling the interests/goals involved in the interaction between agents*;
- The possibility to track the trail of trustee's behaviours (Warnier et al. 2007): *profiling the agents by means of their behavioural patterns*.

Thus, the trust decision is not only based on cognitive predictions but also on the possibility of correcting them. The web of trust is founded on a *circular informational causality*, that is, causality based on shared corrected information that measures the quantity of noise within the communication between the parties of a trust relation. Even if party trust is not strictly a system of control, it can achieve control by means of *information feedbacks*: a negative feedbacks corrects an incorrect prediction; a positive feedback supports the process of communication, even when autonomous agents are concerned by this messaging process: "Clearly the iteration of this process can lead to cycles of positive or negative feedback for each agent, which leads to either a global low or high trust regime" (Ghanea-Hercock 2007).

4 A Revised Notion of Trust: Between Uncertainty and Control

From what I have said so far it emerges that trusting is a complex act, since it brings together what is subject to evaluation and calculation (trustworthiness) and what remains beyond control (the otherness of someone else's behaviour): trusting never involves choice between (a certain amount of) control and (a certain amount of) lack of control, but always involves both of them at the same time. Trust builds a bridge over these two different cognitive states and relations (i.e. trustworthiness and otherness).

Trust contains a sort of "double bind", in the sense of Bateson (Bateson et al. 1956; Bateson 1972) or Derrida (1994). A double bind consists of communication that displays statements that appear to be, to some extent, contradictory. In this case, the double bind (between interacting parties) is represented by a twofold relation: (a) a scalable relation with the agent (i.e. with their qualities, defects, virtues, and reputation) and (b) a non-scalable relation with the otherness of the agent's actual willingness and ability to perform the task. This paradox explains why it is necessary

to provide a rational basis (a reduction of uncertainty), in the form of trust, for deciding to enter into a relationship with someone else: a resolution that is not only concerned with the trust attitude (the cognitive evaluation and calculation) but also with the trust decision (the cognitive lack of evaluation and calculation, which is how we represent what escapes from our own control).

If we look at trust from this perspective, the concept of trust may vary sensibly. Trust is usually conceived only as a relation between parties (agents), whose status is investigated, in order to understand whether or not it is worth trusting. According to our view, trust entails a more complex relation: it is (a) a relation with what can be subject to cognitive evaluation and calculation, on one hand, and (b) a relation with what is subject to a cognitive lack of evaluation and calculation, on the other.

More precisely: *trust is a twofold relation between two relations: with control on one hand, and lack of control on the other, at the same time*. Trust allows us to hold together (a) a representation of (a certain) control and (b) a representation of (a certain) lack of control. The first form of representation (a) is the representation of what appears to us to some extent *uncertain* about someone else's attitudes and behaviours, because it is subject to our limited experience and finite judgment: in this case, someone else's attitudes and behaviours are subject to a form of uncertainty that can be represented, for instance, in the form of *probability, possibility, plausibility measures* or *ranking functions* (Halpern 2005). The second form of representation (b) is the representation of what appears *to us* uncertain about someone else's attitudes and behaviours in a more radical sense, because it is subject to someone else's otherness (i.e. what can evolve without being entirely reduced to our own predictions) and to our non-scalable evaluations. The first form of representation is an investigation and representation of the other (the trustee), which is subject to the cognitive limits of our own calculation, evaluation, judgment and experience. The second form of representation is an investigation and representation of ourselves (the trustor), which is subject to the cognitive limits of our own reflexivity (we progressively acknowledge and try to represent the impact of the uncertain, the unpredictable, and the improbable on our lives (see Taleb 2007).

The relation between cognitive control and lack of control is inherent to trust (namely to the AoT) because it is already intrinsic to human action, if the latter is considered not only as a mere repetition (an instantiation of an internal will) but as what may introduce something new (Arendt 1958). What is new is, by definition, not fully predictable; but only what is new truly nourishes information: "In order to be informative, information must be able to add a distinction and confer something new on what is already known. In this respect, the value of information, what may be called its informativeness, is indeed a function of the kind of 'news' it is capable of conveying, and 'news' differs substantially with respect to what it adds to that which is already known. As a rule, the value of 'news' is traceable to its unique (contingency) and novel (time) character" (Kallinikos 2006, pp. 53–54). That is the raison why trust moves from (incomplete) information towards (new) information.

From this standpoint, trust can no longer be viewed only as a relation between parties but, primarily, as a relation between different (or contrasting) cognitive and informational conditions (control and lack of control), which seem to be able to be associated with any kind of agent. The essential character of trust is to be found in the attempt of *holding together* what is divergent or contrasting. If we lose sight of

this 'bind' or 'holding together', we fail to recognise the phenomenon of trusting. These cognitive and informational conditions (control and lack of control) diverge or contrast not because they are opposite but because they are different by nature: the former can be subject to measurement or calculation, whereas the latter cannot be measured or calculated in the same epistemic terms (which does not mean that it cannot be represented: as we have said, it is first of all a form of reflexivity, an appraisal of our cognitive mind-set). Such a divergence, that is an epistemic difference, does not prevent us to hold together those conditions: this is precisely what trust is for.

Let us take a very straightforward example (for a similar example, see Nissenbaum 2004, p. 176). A mother is doubtful whether or not to delegate a task to her beloved daughter. On the one hand, she would like not to delegate the task to her because she fears that she might be unable to carry it out (*evaluation*); on the other, she would like to delegate her it because of the affection she feels for her (*reflexivity*). Her choice is complicated, at the same time, by judgment and affection that diverge from each other, not because opposite but because they are different by nature: judgment is scalable whereas affection is not. The mother's potential decision to entrust her daughter (which is different from simply delegate her or refuse to delegate her a task) is not a choice between two alternatives (affection over judgment): on the contrary, it is an attempt of *holding together* both alternatives by means of a trust relation. In such relation, in fact, the cognitive process of evaluation is included in the TA towards the daughter's trustworthiness, since the final judgment on her reliability is not discarded but only postponed.

Our perspective about trust appears to be less anthropocentrically biased or, to speak more properly, it has the epistemic advantage that all the entities (agents), which populate the scene of trust are not necessarily to be analysed in anthropocentric terms. We have tried to analyse the process of trusting in a non-anthropocentric manner, at a different level of abstraction. To do so, we have used the notation of a relation between control and lack of control that obeys to a principle of uncertainty. This suggests that shifting the level of abstraction—from the cognitive representation of the interaction between agents to the cognitive representation of the relation between control and lack of control—may well change our observations of agents' behaviours and interactions and hence change the conclusions to be drawn.

5 Whether or Not E-Trust Applies to HA Relations

Does the socio-cognitive TA apply to the case of the relation HA? To answer to this question, we have to analyse the main features of autonomous agents: such analysis can enable us to assess to what extent autonomous agent may be held compatible with socio-cognitive tenets from the point of both the internal attribution and the external attribution of trust (displayed in Section 3.1). This is particularly important as researchers, who suggest that trust is to be reserved for the case of people only (that people can trust other people and not inanimate objects), insist on the following: "One reason to reserve the concept of trust for a relation between people is the role motives and intentions play in it. [...] The philosopher Lawrence Becker, for example, has argued that far more relevant to our readiness to trust are the

motives and intentions we perceive others to have than actions and outcomes" (Camp et al. 2001, p. 6). However, such a consideration is not already a reason to exclude autonomous agents from the realm of trust. I may be willing to endorse the view that: "If it is not possible to design a computer security system without assumptions about human behaviour the design of computer security systems should be informed by philosophical and social sciences theories about trust" (Camp et al. 2001, p. 8). On the other hand, I must also note that to revise cognitive models on the basis of autonomous agents' way of behaving may help us to draw (new) assumptions about human behaviour. In other words, I do not suggest comparing autonomous agents with human agents or to subsume the former within the comprehension of the latter. This could be charged with an accusation of anthropocentrism. I suggest understanding both human and autonomous agents' decision to trust, at a higher level of abstraction, in terms of cognitive states and relations.

5.1 Main Features of Autonomous Agents

The following are the main features inherent to the internal status of the autonomous agents (Franklin and Graesser 1996; Nwana 1996), which are related to the internal attribution of trust:

- Autonomy: the agent's ability to perform a task that originates from the agent itself and is neither a simple adjustment determined by the environment nor the reaction to the intervention of a human user or of another agent; it is the ability to choose between alternatives that are not fully predictable at a determined LoA, which is not that of programmers but of users (Floridi 2004). Compare this notion of agent autonomy with Floridi's definition, given in Section 3.2.
- Persistence: the agent's ability to keep its own internal state while performing its task as well as after its performance.
- Vitality: the agent's ability to face and solve anomalous situations that are otherwise capable of determining in the agent a state of instability that would menace the agent's persistence.
- Organisation: the agent's ability to organise its task, by distributing the work and communicating it to each part that compose its applications;
- Proactivity: the agent's ability to create new situations in the environment, by requesting the intervention of other agents or coordinating their activities in a way to perform its task.

The following are the main features inherent to the autonomous agent's ability to interact with the environment, with users or other agents (Franklin and Graesser 1996; Nwana 1996), which are related to the external attribution of trust:

- Reactivity: the agent's ability to interact with the environment or agents, by changing its states (compare this definition of agent's reactivity with Floridi's more sophisticated explanation of adaptability given in Section 3.2).
- Ability to communicate: the agent's ability to communicate with other agents to cooperate or share resources or learn from experience (this ability is higher when autonomous agents communicate with artificial agents rather than with human agents).

- Mobility: the agent's ability to move within the web or in another intelligent ambient, in order to change its states or partners, when it is necessary to react to the varied environmental conditions.
- Benevolence: the agent's attitude which motivated it to do everything that is required to perform the task exactly, without being deterred from its course of action by external environmental stimulus;
- Attitude to be veridical: the agent's attitude to release true information when interacting with other agents.

These characteristics appear to be fairly compatible with the socio-cognitive analysis of trustor's internal and external attribution of trust. The trustor is enabled to judge both the internal and the external qualities and defects of the autonomous agent, by assigning to each feature:

- An evaluation (also quantitative: <0, 1>) of its probability (possibility, plausibility or ranking);
- An evaluation (also quantitative: <0, 1>) of its relative importance, compared with other features, in relation to a specific goal to be obtained. This ranking function is of great importance to us, since in a cognitive model we judge not only what are the qualities and defects of the trustee; but how significantly these qualities and defect interplay in a given context;
- An evaluation (also quantitative: <0, 1> of its ascertainability by the trustor. This ranking function (how much the trustor can control the trustee and ascertain her cognitive states) is related both to the trustee's ascertainability and correlatively to the trustor's reflexivity (capacity of self-evaluating her own cognitive mind-set).

We could certainly strengthen such results by taking into account elements that directly refer to trustworthiness such as probabilities to defect or to cooperate (Ghanea-Hercock 2007) or rating functions of positive outcomes (Taddeo 2010) but these features can hardly be associated with evaluations, motives, intentions and cognitive judgments. Even if the characteristics we have considered above as associated to the internal and external attribution of trust appear, in their whole, consistent with cognitive schemas and judgments, they may raise problems as to their full applicability to the case of autonomous agents.

5.2 Problems of Applicability of Cognitive Evaluations of TA and TD to AAs

Autonomy is a very troublesome feature for two different kinds of reasons (we will insist on this in the subsection 5.3.2). First of all, it is debatable at what level of abstraction we should investigate agent's autonomy: either at the LoA of programmers (who design, implement and deploy an AA), a group we will call the AA's *developers* (Grodzinsky et al. 2010)" or at the LoA of users? The answer, as we will see later on, may change radically the way we perceive the interaction with autonomous agents: "if computers are perceived as elements of a single undifferentiated network, then trust in computers will increase as computing experience increases" (Camp et al. 2001, p. 1). This question is inherent to the second point at play: that is, the level of predictability of the trusted agent's

behaviour. As it has been stressed (Grodzinsky et al. 2010): "predictability is a central theme that we wish to emphasize", since it "has important consequences in issues of trust". These issues concern not only agents' autonomy but also persistence/change and adaptation (reactivity) and, in the end, the "rationality" itself of the autonomous agent and hence of the AoT. I will consider predictability, firstly, from the point of view of changes driven by autonomy and, then, from the point of view of changes driven by adaptation.

"Predictability is an important attribute from which to draw important distinctions between humans and AAs. AAs are distinct in the sense that we expect that they are capable of much faster changes than humans. Also, the discrete nature of binary encoded programmes increases the likelihood of abrupt and dramatic changes; we expect slower, more gradual changes in processes that at least appear to follow laws described with continuous values and mathematics. (That is, in general we expect binary processes to appear more 'jumpy' and analogue processes to appear 'smoother'). Because software moves at speeds that are beyond the perception of humans, AAs can go through a dramatic self-modification process multiple times during a relatively slow interaction with a human. This sort of change can be disruptive to any existing trust relationship that relies on predictability and that grows out of past experience with that AA" (Grodzinsky et al. 2010). Clearly, it is reasonable to assess that the more predictable a given behaviour is the more certain and stable are the expectations of such behaviour. However, this assessment cannot be turned into an absolute value, since a full predictability would transform incomplete information into a form of certainty that would exclude the interaction between agents from the realm of trust (and place it into the domain of a mere reliance). So, the requirement of predictability is not to be inflated (driven to an absolute value) if one intends to remain in the conceptual area of trust In other words, mere reliance, based on full predictability, is not sufficient in a multi-agent system, where agents continually adapt to their environment—an environment that consists largely of other agents.

As regards to change and adaptation, in addition to what previously said, I wonder whether this ability to change their internal states might at times enable the agents to face unexpected adverse conditions. This may prove true when adaptation is concerned only with changing "the value of a variable". On the contrary, when adaptation is affected with a "self-modifying code", this becomes much more unclear, and we can hardly distinguish whether or not we still interact with the same agent (Grodzinsky et al. 2010). In other words, online users cannot entirely determine themselves on the basis of predictions of future, if such predictions derive from established past knowledge (norms, programmes, instructions, data) that depend on original conditions, which are likely to be, more or less apparently, modified by the mobility and reactivity of other agents. Rather, a rational expectation may be based on present shared information coming from party relations that can be cognitively evaluated and revised (Durante 2008).

In human to human relations, the trustees know that their self-representations and declarations will be judged, relied on and shared: this can deter the agent from a misrepresentation or a false declaration, setting aside the cases in which the trustees take advantage from modifying the original conditions of party relation. Trustees will communicate fairly when they share a concern for the task to be performed and

when their performance is subject to shared information or, to put it differently, when their performance is visible and traceable: it can become a trailed part of agents' past history that trustors can keep track of. Similarly, in human to autonomous agents relations, the issue of "transparency at best, and traceability at least, is a theme" of a great impact, since "if humans are to trust AAs, then AA developers should produce systems whose criteria and process for making decisions are accessible to humans. If these systems' decision-making processes are obscure or hidden, humans are less likely to trust AAs over the long run, and we assert that humans *should* not trust such systems" (Grodzinsky et al. 2010). As regards to mobile agents in particular, the issue of traceability of migration path is crucial and can be faced efficiently by means of distributed trust: i.e. by distributing trust to several hosts on a migration path, which store and communicate information each other, since "hosts cannot cut out or replace the tail of a migration chain (including cycles), as the next receiving host will check the migration with the previous host in the migration chain, using both the host and the sequence number. The essence of this approach is that the responsibility of a migration is spread over two hosts" (Warnier et al. 2007).

We can rephrase here, as regards to H>AAs relations, what we have stated about H>H relations: to the extent to which decisions and behaviours may be traced and expectations be shared, which can give rise to forms of cooperation. In fact, distributed trust turns an individual risk in a collective one, and allows human and autonomous agents to cope more efficiently with the lack of certainties (Durante 2008). It is different to face the lack of certainties by means of a communication process (that does not exclude the presence of an inherent risk but only transforms it into a distributed one) or by achieving a full predictability through design (that aims at eliminating the presence of an inherent risk).

5.3 How to Discuss Arguments Contrary to the Application of Trust to AAs

We have tried to show, to this point, that a socio-cognitive evaluation of AAs features and attitudes is feasible and can enable us to apply both a TA and a TD to the functioning of AAs, while admitting that such application is not devoid of problems. These problems appear at any time we attempt to compare human behaviour (decisions and actions) with autonomous agents' way of behaving: such a comparison leads to a specific problem (Taddeo 2010): "how an artificial agent could be programmed to behave in a manner similar to how humans behave when they report that they have learned to trust someone or something" (Grodzinsky et al. 2010). Of course, this approach, as has been stated, is theoretically correct but too difficult to achieve in practice. Developers take crucial indications from the comparative analysis of human behaviour but they cannot simply attempt to design AAs on the basis of such similarities. They can only attempt to (1) *elaborate* cognitive models to represent agents' behaviour; (2) to figure out the representation of agents' behaviour not as a whole but as *a sum of parts*; (3) to represent each part in *syntactical terms* that make this model applicable to the design of AAs along these guidelines:

- Traceability is to the decision-making process what is memory to human experience (Ferraris 2009): we should not try to fully understand and reproduce

what means to make experience; we have to consider experience as a selection of meaningful data (meaningful because selected);
- Information feedback is to the decision-making process what past experience is to human decision: a trusted third party that registers positive outcomes and makes them part of a distributed trust system may be equivalent to a crucial aspect of human decisions: that is, it is reasonable to expect that, *ceteris paribus*, same information would lead agents to the same decisions;
- Assigning variables to values and weighing values against each other within a frame is to the decision-making process what judging priorities is to human evaluation: evaluation is one of the most complex human activities, since it always entails (from a syntactical standpoint) a combination of (1) variables of values, (2) hierarchies of values, and a (3) framework for hierarchies;
- Relations between states of affairs (conceived in terms of degrees of control) are to the decision-making process what relations between humans are to the construction of human knowledge: like Foucault has shown it, to know is to know what I can control (Foucault 1980, 1994).

These guidelines may help us to discuss the arguments commonly addressed to the possibility to apply trust to AAs with a milder anthropocentric bias (on this, see what has already been said in Section 4).

5.3.1 Is Trust for Human and Not for Artificial Agents?

Trust is a concept entangled with the moral standing of agents: it requires free will, autonomy, reward, appraisal, responsibility, blameworthiness, repentance, betrayal, disapproval, that is, the whole catalogue of moral categories. It is a commonplace that only human beings can endorse moral values, judgments and behaviours. Luciano Floridi's ethics of information (2001, 2004) stands against such commonplace and his moral theory has already been proved to be fruitfully applicable to the domain of trust (Taddeo 2009, 2010). Floridi's ethics of information (2004) is concerned with rebutting five main objections with regards to the possibility of qualifying AAs in moral terms (the teleological objection; the intentional objection; the freedom objection; the responsibility objection and the objection of concreteness), which we cannot expound here (see on this debate, Ethics and Information Technology 2008, pp. 10.2–3). Whether or not Floridi is persuasive—and I think he is to a large extent—there is one point which appears crucial to me in relation to the notion of trust, that is, Floridi's idea of a practical counterfactual that points out the limits of determinism (which I will focus on in the next subsection): "The AAs are already free in the sense of being non-deterministic systems. This much is uncontroversial, scientifically sound and can be guaranteed about human beings as well [...]. All one needs to do is to realise that the agents in question satisfy the usual practical counterfactual: they could have acted differently had they chosen differently, and they could have chosen differently because they are interactive, informed, autonomous and adaptive" (Floridi 2004, p. 17).

As stated above, this idea of a practical counterfactual that makes visible the limits of determinism further enlightens the notion of trust. I have already emphasized this point (Section 4): trust is concerned with a particular form of

reduction of uncertainty. Helen Nissenbaum has expressed it efficaciously, with reference to Niklas Luhmann (1979): "To express the trade-off in Luhmann's terms, we may say that while both trust and security are mechanisms for reducing complexity and making life more manageable, trust enables people to act in a richly complex world, whereas security reduces the richness and complexity" (p. 179).

What I only would add to such formulation is that, thanks to that practical counterfactual (which shows otherness: things could have been happened otherwise), trust enables people to see (and to some extent to represent and to measure) how rich a complex world is: the whole act of trusting is not only a way of entering into the relation with the other, but also of discovering *who* is the other (Durante 2008).

5.3.2 Lack of Trustee's Autonomy

We have already spoken about autonomy but we have to insist here on one core point. As has been noted, the requirement of autonomy is essential, since trust demands the agents to be able to choose between alternatives in a way that is not entirely predictable (like for many intentional states). What we have to highlight is that predictability is not to be judged in some absolute terms but only in relation to the LoA at which agents are studied and analysed: "If a piece of software that exhibits machine learning is studied at a LoA which registers its interactions with its environment, then the software will appear interactive, autonomous and adaptive, i.e. to be an agent. But if the programme code is revealed then software is shown to be simply following rules and hence not to be adaptive. Those two LoAs are at variance. One reflects the 'open source' view of software: the user has access to the code. The other reflects the commercial view that, although the user has bought the software and can use it at will, he has no access to the code. At stake is whether or not the software forms an (artificial) agent" (Floridi 2004, 13).

Obviously, such a consideration does not go so far to exempt developers from responsibility as regards to the design of AAs, both for Floridi (who casts, however, some doubts on this approach: "Our insistence on dealing directly with an agent rather than seeking its 'creator' [...] has led to a nonstandard but perfectly workable conclusion" (Floridi 2004, p. 24))—and for Grodzinsky, Miller &Wolf: "All this does not mean that the concept of 'responsibility' is redundant. On the contrary, our previous analysis makes clear the need for further analysis of the concept of responsibility itself, when the latter refers to the ontological commitment of creators of new AAs and environments. As we have argued, Information Ethics is an ethics addressed not just to 'users' of the world but also to demiurges who are 'divinely' responsible for its creation and well-being" (Floridi 2004, p. 26).

"We prefer that AAs be boringly predictable. We are far more concerned about the trustworthiness of AAs, and far less concerned that they mimic human's adaptability. In almost all situations (with the possible exception of computer gaming), we think that AA developers have a duty to the safety of the public that should restrict their use of the self-modifying code to implement AAs, including limitations on the use of neural nets in AAs" (Grodzinsky et al. 2010).

5.3.3 Trustee's Lack of Repentance or Perception of Trustor's Disapproval

In principle, it may be held true that: "In terms of trust and forgiveness in the context of computer-mediated activities, there is no significant systematic difference in people's reactions to betrayals that originate from human actions, on the one hand, and computer failure, on the other" (Camp et al. 2001, p. 5). Theoretically, trustor's reactions should not vary in relation to the same betrayals of trust. In practice, from a cognitive standpoint, such reaction is likely to be intertwined with a judgment—in particular when the issue of forgiveness is at play—that is concerned with trustee's repentance for betrayal or at least with trustee's perception of the trustor's disapproval for the failure (awareness). From this perspective, a trustee's lack of repentance or perception of trustor's disapproval for betrayal could be an argument against the application of trust to AAs. More precisely, it can be an argument that undermines the possibility to develop an *atmosphere* of trust with AAs on the basis of information feedbacks (on this Section 3.3). On the contrary, we could say that it is precisely an information feedback cycle that could fill the gap between a trustee's lack of perception and repentance: in plain terms, perception of someone else's disapproval and repentance amount to public admission of our own failures and betrayals. In order to obtain a public admission, it is sufficient that failures and betrayals are traced, communicated and stored by a third trusted party that bring them into the general notice of interacting parties.

5.3.4 Trustee's Lack of a 'Fear of Sanctions' for Betrayal

In a previous paper (Durante 2008), I have argued that a trust relation belongs to the area of normativity because social norms grow out of the communication process between trusting parties. If this holds, the normative communication process can be reinforced by the provision of sanctions for betrayal (the same can be said for rewards in case of agents fulfilling their task (Nissenbaum 2004, p. 161). A sanction is not intended here in psychological terms as the fear for the trustor's reaction or judgment. If a sanction is conceived in such terms, this leads us to understand fear as an emotion that only human beings can experience. Conversely, a sanction can be described in strict legal terms as a negative consequence that stems from the betrayal (the unobserved norm). How can a consequence be negative? It can be negative in that it deprives me of something. In other words, a sanction is a diminishment of me. If a sanction is interpreted in this way, fear should no longer be conceived as a human emotion: it can be intended as a scalable state of affairs that correspond to a diminishment that lowers my probabilities to act. Probabilities to act, attached to a negative success rate, can be lowered down until the point that the trustee will be automatically prevented from acting by a further diminishment in its success rate: in this perspective, we may say that a 'fear of sanctions' may prevent it from acting badly (on the applicability of sanctions to artificial agents see L. Floridi 2004).

5.3.5 Lack of Embodied Relations Between Agents

The last point is to reiterate that human touch is not, in my view, a needed requirement for trust. From a philosophical point of view, it would take time to

expound this reflection: suffice to say that human relations are not only empirical relations that require touch, contact, physical proximity etc. The fear for a progressive disembodiment of human relations and experiences is not caused by the fact that embodiment is a necessary character of both relation and experience. On the contrary, it originates from the fact that embodiment has became, throughout time, a sort of guarantee of the truthfulness of relations and experiences. An empirical, embodied situation seems to endow us with the feeling and the epistemic belief of being able to have control over our relations and experiences. This brings us back to problem of control—the real cognitive and practical problem—which we have started from. Things are far more complicated than we have attempted to represent, since it is an oversimplification to speak of a relation between control and lack of control, which is only justified by the analysis of trust. A (certain) lack of control is, most of the times, insular to any cognitive and epistemic 'control' (as I have defined it from the start): it is not something that comes from the outside of a situation of control but it is inherent to it. Even when I look at my clock to check what time it is, I can easily 'control' what time it is, but I can still doubt whether or not my clock is on time. I always need, in the case, a further guarantee, which is not necessarily at hand. In this perspective, trust is concerned with the agents' need (be human or artificial agents) to step out of their limited circle of guaranties (of course, the reverse situation is not necessarily true: not all that escapes from our control deserves trust). As a final point, trust is not only a relation between interacting agents but, primarily, a cognitive relation with what remains out of control *within* what we believe to hold control over.

References

Arendt, H. (1958). *The human condition*. Chicago: Chicago University Press.

Bateson, G. (1972). *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology.* Chicago: Chicago University Press.

Bateson, G., Jackson, D. D., Haley, J., & Weakland, J. (1956). Toward a theory of schizophrenia. *Behavioural Science, 1*, 251–264.

Camp, J., McGrath, C., & Nissenbaum, H. (2001). Trust: A collision of paradigms. In *Financial Cryptography 2001: Conference Proceedings*, Springer Verlag Lecture Notes in Computer Science. Available at: http://www.nyu.edu/projects/nissenbaum/main_cv.html (accessed 15 June 2010).

Castelfranchi, C., & Falcone, R. (2007). "Trust Theory". available at: http://www.istc.cnr.it/T3/trust (accessed 15 June 2010).

Castelfranchi, C., & Falcone, R. (2008). "Socio-cognitive model of trust: Basic ingredients". Available at: http://www.istc.cnr.it/T3/trust (accessed 15 June 2010).

Chiodi, G. M. (2000). *Equità. La regola costitutiva del diritto*. Torino: Giappichelli.

Derrida, J. (1994). *Given time: I. Counterfeit Money.* Chicago: Chicago University Press. Translated by P. Kamuf.

Durante, M. (2008). What model of trust for networked cooperation? Online social trust in the production of common goods (Knowledge Sharing). In: T. W. Bynum, M. Calzarossa, I. De Lotto, & S. Rogerson (Eds.), Living, working and learning beyond technology. Proceedings of the Tenth International Conference Ethicomp 2008, Mantova, Tipografia Commerciale, pp. 211–223.

Ethics and Information Technology (2008). In: C. Ess (Ed.), Luciano Floridi's Philosophy of Information and Information Ethics: Critical Reflections and the State of the Art. 10. pp. 2–3.
Ferraris, M. (2009). *Documentalità. Perché è necessario lasciar tracce*. Roma-Bari: Laterza.
Floridi, L. (2001). Artificial evil and the foundation of computer ethics. *Ethics and Information Technology, 3*(1), 55–66.
Floridi, L. (2004). On the morality of artificial agents. *Mind and Machine, 14*(3), 349–379.
Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329.
Foucault, M. (1980). In C. Gordon (Ed.), *Power/Knowledge: Selected interviews and other writings, 1972–1977*. New York: Pantheon.
Foucault, M. (1994). *The order of things: An archaeology of the human sciences*. New York: Vintage.
Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program? A taxonomy for autonomous agents. In: Proceedings of the Third International Workshop on Agent Theories, Architectures and Languages. Springer-Verlag. Available at: http://www.msci.memphis.edu/franklin/AgentProg.html (accessed 15 June 2010).
Ghanea-Hercock, R. (2007). Dynamic trust formation in multi-agent system. In Tenth international workshop on trust in agent societies at the autonomous agents and multi-agent systems conference (AAMAS 2007), Hawaii; May 15, 2007. Available at: http://www.istc.cnr.it/T3/trust (accessed 15 June 2010).
Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2010). Developing artificial agents worthy of trust: 'Would you buy a used car from this artificial agent?'. In: Proceedings of CEPE, June 2009, Greece. pp. 26–28
Halpern, J. (2005). *Reasoning about uncertainty.* Cambridge: MIT Press.
Kallinikos, J. (2006). *The consequences of information. Institutional implications of technological change*. Cheltenham: Edward Elgar.
Luhmann, N. (1979). Trust: a mechanism for the reduction of social complexity. In N. Luhmann (Ed.), *Trust and power: Two works* (pp. 1–103). New York: Wiley.
Moor, J. (1985). What is computer ethics? In T. Ward Bynum (Ed.), *Computers & ethics* (pp. 266–275). Malden: Blackwell.
Nissenbaum, H. (2004). Will security enhance trust online, or supplant it? In R. M. Kramer & K. S. Cook (Eds.), *Trust and distrust in organizations: Dilemmas and approaches* (pp. 155–188). New York: Sage.
Nwana, H. (1996). Software agents: an overview. *Knowledge Engineering Review, 11*(3), 1–40.
Resta, E. (2009). *Le regole della fiducia*. Roma-Bari: Laterza.
Scillitani, L. (2007). *Fiducia, diritto, politica. Prospettive antropologico-filosofiche*. Torino: Giappichelli.
Taddeo, M. (2009). Defining trust and e-trust: from old theories to new problems. *International Journal of Technology and Human Interaction, 5*(2), 23–35.
Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and Machines, 20*(2), 243–257.
Taleb, N. N. (2007). The Black Swan. In *The impact of the highly improbable*. New York: Random House.
Warnier, M., Oey, M., Timmer, R., & Brazier, F. (2007). Secure migration of mobile agents based on distributed trust. In R. Falcone, S. Barber, J. Saboteur-Mir, & M. Singh (Eds.), *Trust in agent societies* (pp. 112–116). online at http://www.istc.cnr.it/T3/trust.