Modelling trust in artificial agents, a first step toward the analysis of e-trust

Mariarosaria Taddeo

Information Ethics Group, University of Oxford, UK,

mariarosariataddeo@gmail.com

Modelling trust in artificial agents, a first step toward the analysis of e-trust

Abstract

This paper provides a new analysis of *e-trust*, trust occurring in digital contexts, among the artificial agents of a distributed artificial system. The analysis endorses a non-psychological approach and rests on a Kantian regulative ideal of a rational agent, able to choose the best option for itself, given a specific scenario and a goal to achieve.

The paper first introduces e-trust describing its relevance for the contemporary society and then presents a new theoretical analysis of this phenomenon. The analysis first focuses on an agent's *trustworthiness*, this one is presented as the necessary requirement for e-trust to occur. Then, a new definition of e-trust as a second-order-property of first-order relations is presented. It is shown that the second-order-property of e-trust has the effect of minimising an agent's effort and commitment in the achievement of a given goal. On this basis, a method is provided for the objective assessment of the levels of e-trust occurring among the artificial agents of a distributed artificial system.

Keywords: artificial agent, artificial distributed system, e-trust, trust, trustworthiness.

1. Introduction

Trust is a widely diffused phenomenon, affecting many of our daily practises. It is the key to effective communication, interaction and cooperation in any kind of distributed system, including our society (Lagenspetz 1992). As Luhmann (Luhmann 1979) suggests, "a complete absence of trust would prevent even getting up in the morning", (p. 4). It is because we trust the other parts of the system to work properly that we can delegate some of our tasks and may focus only on the activities that we prefer. As in Plato's *Republic*, for example, one can trust other agents to defend the city and dedicate one's time to studying philosophy.

Although trust is largely recognised as an important issue in many fields of research, we still lack a satisfactory analysis. Moreover, in recent years, with the emergence of trust in digital contexts – known as *e-trust* – new theoretical problems have arisen.

The debate focuses on two main issues: the *definition* and the *management* of trust and e-trust. While the definition of trust and e-trust is left to sociologists and philosophers, the management of their occurrences is a topic of research for ICT.

In digital contexts, like e-Bay or an artificial distributed system, all the interactions are mediated by digital devices and there is no space for direct and physical contacts. In the digital sphere moral and social pressures are also perceived differently than in real environments. These differences from the traditional context in which trust occurs give the raise to a major problem, that is "whether trust is affected by environmental features in such a way that it can only occur in non-digital environments, or it is mainly affected by features of the agents and their abilities, so that trust is viable even in digital contexts", (Taddeo 2009) (p. 29). Those who argue against the case of e-trust, see Nissenbaum (Nissenbaum 2001) for example, suggest that the absence of physical interactions and of moral and social pressure in the digital context constitute an obstacle to the occurrence of trust, and that for this reason e-trust should not be considered as an occurrence of trust in digital contexts but as a different phenomenon. Despite their plausibility, the objections against e-trust can be rebutted (Taddeo 2009). Several accounts of e-trust consider this phenomenon as an occurrence of trust in digital environments, see for example (Weckert 2005), (Vries 2006), (Seamons, Winslett et al. 2003), and (Papadopoulou 2007).

Trust and e-trust have been defined in different ways: as a probabilistic evaluation of trustworthiness in (Gambetta 1998) and (Castelfranchi and Falcone 1998) as a relationship based on ethical norms (Tuomela and Hofmann 2003), and as an agent's attitude in (Weckert 2005). Unfortunately, all these analyses focus only on the trustor's beliefs, and so provide a partial explanation of the phenomenon, leaving many questions unanswered. These include, for example, what the effects of trust and e-trust on the involved agents' behaviours are, for what reason an agent decides to trust, or the role of trust and e-trust in the development of social systems.

The research for the management of trust and e-trust seeks to identify the necessary parameters for their emergence and to define a method for the objective assessment of the level of trust and e-trust occurring between two or more agents. This research often rests on theoretical analyses of trust and e-trust uncritically, thus inheriting their limits. It will therefore benefit from an analysis of trust and e-trust able to overcome these boundaries and to answer the questions described above.

This paper seeks to contribute to the debate by presenting a new analysis of e-trust. This analysis rests on a Kantian regulative ideal of a rational agent, able to choose the best option for itself, given a specific scenario and a goal to achieve. E-trust is considered a result of a rational choice that is expected to be convenient for the agent. This approach does not reduce the entire phenomenon of trust to a matter of pure rational choice but, in this paper, we shall be concerned with the specific occurrences of e-trust among artificial agents (AAs) of distributed systems. So, given this scenario, it is simply more realistic to consider these AAs to be designed (or at least designable) as fully rational. One might object that trust involves more than rational choice and that a model unable to consider other than rational aspects could not provide a satisfactory analysis of this phenomenon. This is correct and therefore I will show how the provided analysis can be extended to consider non-rational factors as well.

Here is a summary of the paper. In section 2, I describe e-trust in general. In section 3, I analyse its foundation. In section 4, I present and defend a new definition of e-trust. In section 4.1, I describe a method for the objective assessment of the levels of e-trust by looking at the implementation of the new definition in a distributed artificial system. In section 5, I show how the new model can be used to explain more complex occurrences of e-trust, such as those involving human agents (HAs). Section 6 concludes the paper.

2. E-trust

Trusting AAs to perform actions that are usually performed by human agents (HAs) is not science-fiction but a matter of daily experience. There are simple cases, such as that of refrigerators able to shop online autonomously for our food[1], and complex ones, such as that of Chicago's video surveillance network,[2] one of the most advanced in the US. In the latter case, cameras record what happens, and by matching that information with patterns of events such as murder or terrorist attacks, seek to recognise dangerous situations. The crucial difference between this system and the traditional ones is that it does not require a monitoring HA. In this case, HAs – the

[1] http://www.lginternetfamily.co.uk/fridge.asp

[2] http://www.nytimes.com/2004/09/21/national/21cameras.html

entire Chicago police department – trust an artificial system to discern dangerous situations from non-dangerous ones.

In the same way, HAs trust AAs to act in the right way in military contexts. The US army used robots 'Warrior' and 'Sword' in Iran and Afghanistan.[3] More sophisticated machines are now used at the borders between Israel and Palestine in the so-called 'automatic kill zone'.[4] The robots are trusted to detect the presence of potential enemies and to mediate the action of the HAs. It is to be hoped that, in a not too distant future, we shall trust these robots to distinguish military enemies from civilians, and not to fire upon the latter.[5]

Even more futuristic are the cases of AAs that trust each other without the involvement of HAs. Consider, for example, unmanned aerial vehicles like the Predators RQ-1 / MQ-1 / MQ-9 Reaper. These vehicles are "long-endurance, medium-altitude unmanned aircraft system for surveillance and reconnaissance missions".[6] Surveillance imagery from synthetic aperture radar, video cameras and infrared can be distributed in real time to the front line soldier and the operational commander, and worldwide, via satellite communication links. The system in the vehicle receives and records video signals in the ground control station and can pass them to another system, the Trojan Spirit van, for worldwide distribution or directly to operational users via a commercial global broadcast system. In this case, there are two occurrences of e-trust: that between the artificial system and HAs, and that between the artificial systems that manage the data flow. The Predator's system of data-collecting trusts the broadcast systems to acquire the data and transmit them to the users without modifying or damaging them or disclosing them to the wrong users.

As the previous examples illustrate, there are at least two kinds of e-trust. In the first, HAs trust (possibly a combination of) computational artefacts, a digital devices or services, such as a particular website, to achieve a given goal. The users of eBay, for example, trust its rating system. In the second, only AAs are involved. This is the case of trust in distributed artificial systems.

This paper is concerned with e-trust between AAs in a distributed system. This occurrence of e-trust is simpler to describe and model than those in which HAs are

[3] http://blog.wired.com/defense/2007/08/httpwwwnational.html

[4] http://blog.wired.com/defense/2007/06/for_years_and_y.html

[5] http://blog.wired.com/defense/2007/06/for_years_and_y.html

[6] http://www.airforce-technology.com/projects/predator/

involved. This is because trust and e-trust are primarily the consequences of decisions made by the trustor. An analysis of e-trust has to make clear the criteria by which an agent decides to trust, and this is easier to do for AAs than for HAs, because the criteria by which an HA decides to trust are several and heterogeneous (they include economic, attitudinarian and psychological factors), and their relevance varies across different circumstances. AAs are computational systems situated in a specific environment, and able to adapt themselves to changes in it. Since AAs are not endowed with mental states, feelings or emotions, psychological and attitudinarian criteria are irrelevant. In analysing an AA's decision process, one knows exactly the criteria by which it decides. This is because AAs are artefacts whose criteria are either set by the designer or learned following rules set by the designer (Wooldridge 2002).[7] The hope is that, by focusing on such a streamlined and fully-controlled scenario, one might be able to identify the fundamental features of e-trust more easily.

Let us now consider e-trust in more detail. I will begin the analysis of e-trust from its foundation: an AA's trustworthiness.

3. The foundation of e-trust: an objective assessment of trustworthiness

Trust is often understood as a relation that holds when (a) one of the parts (the *trustor*) chooses to rely on another part (the *trustee*) to perform a given action, and (b) this choice rests on the assessment of the trustee's *trustworthiness*. The trustor's assessment of the trustee's trustworthiness is usually considered the foundation of trust and e-trust. It is defined as the set of beliefs that the trustor holds about the potential trustee's abilities, and the probabilities she assigns to those beliefs. This definition needs to be revised (reference removed due to double blind review). Potential trustors do not consider only their subjective beliefs but also objective aspects, such as the results of the performances of the potential trustee. Consider how trustworthiness is assessed in online contexts, for example in on-line purchases. It has been shown - see (Castelfranchi and Falcone 1998) and (Corritore, Kracher et al. 2003) - that the potential seller's trustworthiness is assessed on the basis of well-

[7] AAs are computational systems situated in a specific environment and able to adapt themselves to changes in it. They are also able to interact with the environment and with other agents, both human and artificial, and to act autonomously to achieve their goals**.** AAs are not endowed with mental states, feelings or emotions. For a more in depth analysis of the features of AAs see Floridi, L. and J. Sanders (2004). "On the Morality of Artificial Agents." Minds and Machines **14**(3): 349-379.

defined and objective criteria (such as previous experiences made with the same website, the brand, the technology of the web site, and the seals of approval) that have to be met for users to consider the website trustworthy.

This feature of the trustworthiness' assessment is well represented in the analysis of the model of a distributed system of rational AAs discussed in this paper. In order to determine a potential trustee's trustworthiness, the AAs calculate the ratio of successful actions to total number of actions performed by the potential trustee to achieve a similar goal. Once determined, this value is compared with a threshold value. Only those AAs whose performances have a value above the threshold are considered trustworthy, and so trusted by the other AAs of the system.

The threshold is a parameter that can be either fixed by the designer or determined by the AAs of the system on the basis of the mean performance value of the AAs. Generally speaking, the threshold has a minimum value below which it cannot be moved without imposing a high risk on the trustor. Trustworthiness is then understood as a measure that indicates to the trustor the probability of her gaining by the trustee's performances and, conversely, the risk to her that the trustee will not act as she expects. The minimum value of the threshold is calculated by considering both the risk and the advantage to which a rational AA might be subject in performing a given action by itself. The threshold must be higher than this value, because for a rational AA it is not convenient to consider as trustworthy an AA that can potentially determine a higher risk and an equal advantage to the one that the AA would achieve acting by itself.

Note that the threshold is set higher than the mean value of the AAs' performances. This because the AAs make decisions that, according to their information, will best help them to achieve their goals,[8] and an AA's trustworthiness is a key factor.

[8] These AAs are assumed to comply with the axioms of rational choice theory. The axioms are: (1) completeness: for any pair of alternatives (x and y), the AA either prefers x to y, prefers y to x, or is indifferent between x and y. (2) Transitivity: if an AA prefers x to y and y to z, then it necessarily prefers x to z. If it is indifferent between x and y, and indifferent between y and z, then it is necessarily indifferent between x and z. (3) Priority: the AA will choose the most preferred alternative. If the AA is indifferent between two or more alternatives that are preferred to all others, it will choose one of those alternatives, with the specific choice among them remaining indeterminate.

Trustworthiness is the guarantee required by the trustor that the trustee will act as it is expected to do without any supervision. In an ideal scenario, rational AAs choose to trust only the most trustworthy AAs for the execution of a given task. If threshold values are high, then only very trustworthy AAs will be potential trustees, and the risk to trustors will be low. This is of course a self-reinforcing process.

There are two more points to note. First, the AAs considered in this analysis assess trustworthiness with respect to a specific set of actions. For this reason, trustworthiness is not a general value: it does not indicate the *general reliability* of an AA, but its *dependability* for the achievement of a specific (kind of) goal. It might happen that an AA that is trustworthy for the achievement of a given goal, such as for example collecting the data about the last *n* online purchases of a user, is not trustworthy for another goal, such as guaranteeing the privacy of the same user.

Second, e-trust does not occur *a priori*. A rational AA would not trust another AA until its trustworthiness had been assessed. Given that trustworthiness is calculated on the basis of the results of performed actions, e-trust could not occur until the potential trustees had performed some actions relevant to the matter of the trust in its regards. In systems with only rational agents, no agent whose trustworthiness is not yet measured is trusted. This implies that there might be cases in which the trustee's trustworthiness has not been assessed with respect to the specific matter of trust. In these circumstances the interaction between the two AAs still happens but there is not occurrence of e-trust, because the trustor cannot assess the trustee's trustworthiness. Like Crusoe and Friday in Defoe's book, the two AAs will interact at fist without trusting each other. e-Trust will emerge between the two in the curse of time, on the basis of results of their previous interactions and if, and only if, the outcomes of the trustee's performances indicate a high level of trustworthiness.

The reader should recall that the AAs described in this analysis are purely rational agents, they could take the risk involved in trusting another agent only if the potential trustee can objectively be considered trustworthy. As we shall see in section 5, a less rational agent might follow different criteria in deciding whether to trust another agent.

One might consider the dependence of the occurrence of e-trust from the assessment of the trustee's trustworthiness as a limit, and object that, since the appraisal of trustworthiness presupposes past interactions, it is impossible to explain the occurrence of e-trust in one-shot interactions, which are indeed very common in the occurrence of e-trust in a distributed system.

However, it is important to realize that an agent's trustworthiness is assessed on the agent's past performances and not on its iterated interactions with the same agent. For the assessment of trustworthiness it does not matter whether the trustee performed its actions alone or by interacting with the same trustor, or even another agent. Trustworthiness is like a chronicle of an agent's actions, a report that any agent of the system considers in order to decide whether to trust that agent. Consider the way trustworthiness is assessed in reputation systems, for example in Web of Trust (WOT).[9] WOT is a website reputation rating system. It combines together ratings submitted by individual users about the same trustee to provide community ratings. The resulting collective rating is then shared with all the users who can check it before deciding to trust a particular agent. In this way, any user has the possibility to assess an agent's trustworthiness even in case of one-shot interaction.

On the model I have described, an AA's trustworthiness is a value determined by the outcomes of its performances. The high trustworthiness of an AA guarantees its future good behaviour as a potential trustee and is a justification for the trustor to take the risk of trusting it. Trustworthiness is like the year of foundation written underneath the name of a brand, such as that in Smith & Son 1878. It tells us that since 1878, Smith & Son have being doing good work, satisfying their customers and making profit, and it is offered as a guarantee that the brand will continue to do business in this way. In the same way, an AA's trustworthiness is a testament to its past successes, which is cited as a guarantee of the future results of its actions. This is the starting point for the analysis of e-trust that I will provide in the next section.

4. E-trust: a property of relations

Let me begin by presenting a definition of e-trust that is based on the definition of trustworthiness given in the previous section.

> **Definition**: Assume a set of first-order relations functional to the achievement of a goal and that two AAs are involved in the relations, such that one of them (the trustor) has to achieve the given goal and the other (the trustee) is able to perform some actions in order to

[9] http://en.wikipedia.org/wiki/WOT:_Web_of_Trust

> achieve that goal. If the trustor chooses to achieve its goal by the action performed by the trustee, and if the trustor rationally selects the trustee on the basis of its trustworthiness, then the relation has the property of minimising the trustor's effort and commitment in the achievement of that given goal. Such a property is a second-order property that affects the first-order relations taking place between AAs, and is called *e-trust*.

This definition differs from others provided in the literature for three reasons. First, it defines e-trust as a second-order property of first-order relations, whereas e-trust is classically (if mistakenly) defined as a first-order relation that occurs between agents. Second, the definition stresses not only the decisional aspects of trust, but also the effects of e-trust on the actions of AAs. Third, the definition highlights the fact that e-trust is goal-oriented. These features deserve some clarification.

Let us first focus on the definition of e-trust as a second-order property of first-order relations. Consider for example, a MAS in which AAs interact in commercial transactions.[10] In this case the AAs of the systems are divided in two categories, the sellers and the buyers. The sellers put on the market their products and the buyers have a set of goods that need to purchase. A seller (B) and a buyer (A) start the negotiation process anytime that a seller's offer satisfies the buyer's need. The negotiation process might occur with or without trust, depending on the trustworthiness value associated to the seller.[11] According to the definition, when e-trust occurs, there is a first-order relation, purchasing, which ranges over the two agents, and e-trust, which ranges over the first-order relation and affects the way it occurs. In symbols, T (P (A, B, g)), where T is the property of e-trust, which ranges over the relation P, purchasing, occurring between the agents A and B about some good g. Cases like this are usually explained in the literature by arguing that there are

[10] These systems are widely diffused, there is a plethora of MAS able to perform tasks such as product brokering, merchant brokering and negotiation. Such systems are also able to address problems like security, trust, reputation, law, payment mechanisms, and advertising. Guttman, R., A. Moukas, et al. (1998). "Agent-Mediated Electronic Commerce: A Survey." Knowledge Engineering Review **13**(3): 147-159. Nwana, H., J. Rosenschein, et al. (1998). Agent-Mediated Electronic Commerce: Issues, Challenges and some Viewpoints. Autonomous Agents 98, ACM Press.

[11] The reader may consider this process similar to the one that occurs in e-commerce contexts where HAs are involved, like e-Bay for example.

two first-order relations, purchasing and e-trust, occurring at the same time. This explanation could be used to argue against the definition of e-trust as a second-order property. Let us accept it for a moment. In the case of the example above, we have a relation P (A, B, g) and a relation T (A, B), where both P and T range over the two AAs. In this case, B is trusted as such, and not merely for the performance of a specific action. From the relation T (A,B) it follows that the buyer trusts the seller to be able to perform any sort of tasks, from honestly selling its goods to guarantying the security of the system or to managing the information flaw in the system. This cannot be accepted, it would be like saying that by trusting a seller on e-Bay one also trusts that seller to be a good lawyer or a good astronaut.

One may argue that A generally trusts B to be an honest seller. If this were so, e-trust would range over all the first-order relations of selling that occur between A and B, where B acts, or is supposed to act, honestly. There is a first-order relation between A and B, which occurs every time B acts honestly and which falls in the range of T (A, B). But this simply supports the definition that I provided early in this section.

The analysis of e-trust presented in this paper differs from the one provided in literature - see (Castelfranchi and Falcone 1998), (Weckert 2005), and (Tuomela and Hofmann 2003) - partly because it takes e-trust to be a property of relations. The definition states that e-trust affects relations between AAs, minimising the trustor's effort and commitment in achieving a goal. The following example might help to explain this.

Consider again the case of AAs described above. If the buyer trusts the seller, and the seller is trustworthy, then purchasing occurs more easily than it would have done had the buyer not trusted the seller. The absence of e-trust would force the buyer to spend its time looking for information about the seller and about its reliability, before purchasing something from that seller.

E-trust relations minimise the trustor's effort and commitment in two ways. First, the trustor can avoid performing an action because it can count on the trustee to do it instead. Second, the trustor does not supervise the trustee's performances without taking a high risk about the trustee's defection. This is a peculiarity of e-trust relations: the absence of supervision is justified there by the trustee's trustworthiness, which guarantees that the trustee is able to perform a given action correctly and autonomously.[12]

This minimisation of the trustor's effort and commitment allows the trustor to save time and the energy that it would have spent in performing the action that the trustee executes, or in supervising the trustee.

The level of resource saving is inversely related to the level of e-trust. The higher the e-trust in another AA, the less the trustor has to do in order to achieve its goal and to supervise the trustee. Hence, the more an AA trusts, the less it spends on resources. In this way, e-trust allows the trustor to maximise its gain (by achieving its goal) and minimise its loss of resources.

The inverse relation between the level of e-trust and the level of resource-use falls under the mini-max rule.[13] On the basis of this rule, e-trust and a trustor's resources are variables of the same system, and are correlated in such a way that the growth of one causes the decrease of the other. By using this system and knowing one of the variables, it is possible to determine the other. The reader should recall that AAs' resources, time and energy, can be objectively quantified. This allows for an objective assessment of the level of e-trust affecting a relation. The model for such an assessment is provided in section 4.1.

Before considering in more details the relation between e-trust and the resources of an AA, let us consider the third reason why my definition differs from previous ones. This is that it considers e-trust to be goal-oriented. That e-trust is goal-oriented follows from the teleological orientation of AAs: they act always to achieve some goal, and e-trust is part of a strategy for best achieving those goals. It also follows from the definition in section 3 of trustworthiness as the ability of an AA to achieve a given goal. We can now look at the model.

4.1 An objective model for e-trust levels assessment

The inverse relation between the level of e-trust and the level of resources can be formalised by equating the level of e-trust (y) to the cube of the inverse of the

[12] Note that "action" indicates here any performance of an AA, from, for example, controlling an unmanned vehicle to communicating information or data to another AA. For the role of trust in informative processes see (reference removed for double blind review).

[13] As the reader might already know a mini-max rule is a decision rule used in decision and game theory. The rule is used to maximise the minimum gain, or inversely to minimise the maximum loss.

resources (x): $y = (1 - 2x)^3$. This equation allows one to draw a curve such as the one in Figure 1.

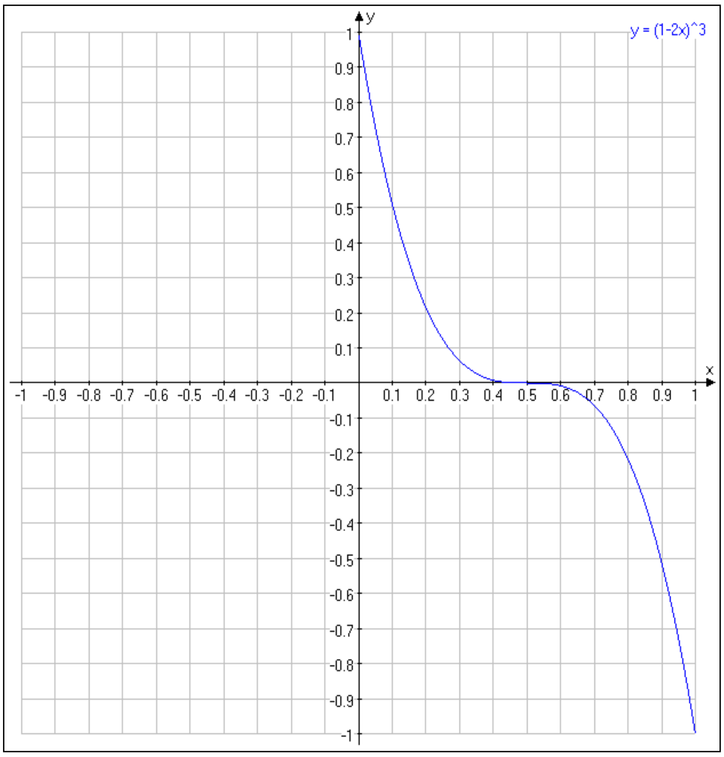

Figure 1 Graph showing the levels of e-trust provided by the equation $y = (1\text{-}2x)^3$

The curve goes from the maximum level of e-trust, 1, achieved when the level of resources used is equal to 0, to the maximum level of distrust, achieved when the level of the resources used is equal to 1. The power has been introduced to account for the impact of other factors on the occurrence of e-trust, such as the context, the needs of the trustor and the possible urgency of performing an action affected by e-trust. The equation can be modified using different odd powers, depending on how much the extra factors are supposed to affect the emergence of e-trust. The graph will always show a curve going from the maximum level of e-trust to the maximum level of e-distrust. But, as the reader can see from the graph in Figure 2, the same level of e-trust corresponds to different levels of resources with different powers. More precisely, the power is an expression of the friendliness of the context. The higher the power, the friendlier the context. In a friendly context, AAs are less exposed to dangers, and the level of alert and of supervision by an AA of other AAs' performances is lower. The level of resources used to supervise the trustee's performances counts for more in assessing the level of e-trust. For a higher power ($p > 3$), the same amount of resources corresponds to a lower level of e-trust.

This model yields both theoretical and practical results. The theoretical results are as follows. The graph shows that there are three macro sets of e-trust levels: ($y > 0$), ($y = 0$), and ($y < 0$). These sets define a taxonomy of e-trust relations allowing for a distinction between e-trust positive relations ($y > 0$), e-trust neutral relations ($y = 0$), and e-trust negative, or e-distrust, relations ($y < 0$).

The presence of e-distrust relation in the taxonomy derived from the model is an important feature, for it shows that the definition defended in this article enables us to explain the entire range of e-trust relations in one consistent framework. E-distrust relations occur when an AA (the dis-trustor) considers the other AA (dis-trustee) untrustworthy and hence uses a high level of its resources to supervise (or ultimately replace) the dis-trustee's performances. Again, as the reader can see from the graph in figure 1, the level of e-trust is lower than 0, and it decreases with the increase of the level of resources used. Hence, e-distrust relations, like e-trust positive and neutral relations, fall under the mini-max rule and are in an inverse relation, as described in section 4.

The practical result is that this method can be implemented by systems for e-trust management in order to determine the level of e-trust occurring in a given relation, and to identify cases of distrust.

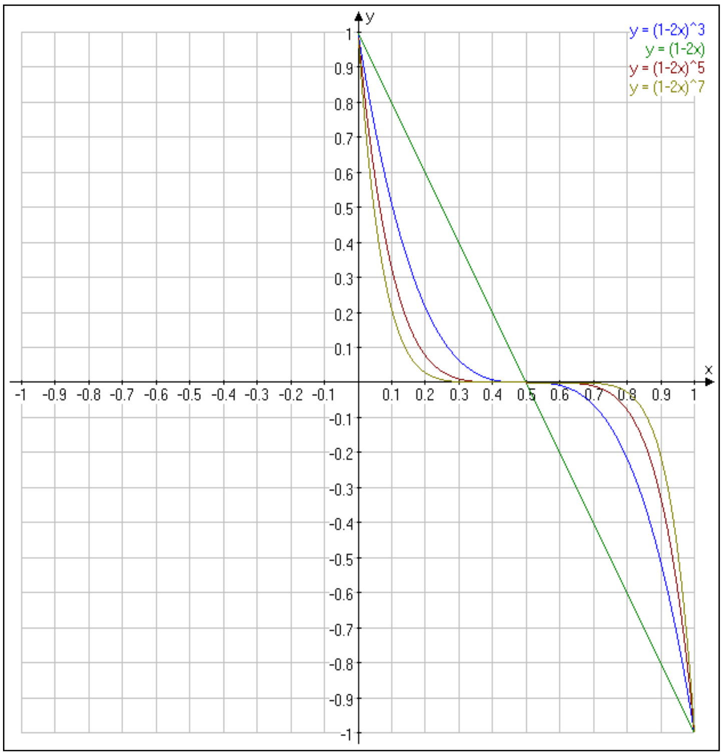

Figure 2 Graph showing the levels of e-trust provided by using the equation y = (1 -2x) with different powers.

5. From e-trust to trust

Now that I have described the definition and model of e-trust in detail, the challenge is to extend them to explain more complex occurrences of trust and e-trust, such as those in which HAs are involved. In this section, I will briefly sketch how the challenge can be met.

When HAs are involved, the conditions for the emergence of trust are more complex. There are various criteria, and they change according to the context and to the agents involved. Consider, for example, the case of a rational HA, R, and of a gullible HA, G. Agent R trusts her bank to pay her bills automatically. Her trust in the bank rests on many criteria, such as information in reviews she read about the bank, the good impression she had during a meeting with one of the bank's employees, and the fact that all her bills have been regularly paid so far. R's goal is to have her bills paid and her account properly managed. Now consider agent G, who also trusts his bank to pay his bills, but whose trust rests on a different assessment of its trustworthiness. G chose a bank indiscriminately, and delegated the bank to pay his bills. G's goal is not to have to care about his bills and about the way the bank manages his account. He did not check that the bank would offer a good service. He prefers to take a high risk than be concerned with the trustworthiness of the trustee. Generally speaking, an agent like G prefers to trust, because it allows him to delegate an action and so not to have to care about it.

In the cases of both R and G, there is a first-order relation between an agent and the bank. The relation occurs in the way determined by the presence of trust, delegation and absence of supervision. What distinguishes the two cases is how the trustworthiness of the trustee (the bank) is assessed, and the kind of advantage that is conferred by trusting. Given their heterogeneity, the criteria for the assessment of the trustworthiness of the bank, and the benefits of trusting cannot be specified *a priori*. They must be specified case by case. This is why it is initially rather difficult to extend the definition of e-trust to more complex cases.

This difficulty can be overcome if we make the definition and the model more abstract. Trust and e-trust are still defined as second-order properties of first-order relations, able to determine some advantages for the trustor and grounded on the assessment of the trustee's trustworthiness, but the criteria for the assessment of trustworthiness and the benefits of trusting are not specified. Their specification is left to the analysis of the specific cases, while the definition provides a general framework able to explain consistently the occurrences of both trust and e-trust.

By using the methodology of the levels of abstraction (LoA)[14] (Floridi 2008), it is possible to set the same definition at different levels of abstraction. From a high LoA, the definition does not take into consideration the criteria for the assessment of trustworthiness, but focuses only on the more general aspects of trustworthiness as the foundation of trust and e-trust. The criteria for assessing trustworthiness are specified at a lower LoA. Shifting from higher to lower LoAs is like looking at a map of a city first with the naked eye, and then with a magnifying glass. In the first case, one sees a

[14] In the Theory of Level of Abstraction (LoA), discrete mathematics is used to specify and analyse the behaviour of information systems. The definition of a LoA is this: given a well-defined set X of values, an observable of type X is a variable whose value ranges over X. A LoA consists of a collection of observables of given types. The LoA is determined by the way in which one chooses to describe, analyse and discuss a system and its context. A LoA consists of a collection of observables, each with a well-defined possible set of values or outcomes. Each LoA makes possible an analysis of the system, the result of which is called a model of the system. Evidently, a system may be described at a range of LoAs and so can have a range of models. More intuitively, a LoA is comparable to an 'interface', which consists of a set of features, the observables.

general model of a city, which would potentially fit any city. This is like choosing a high LoA. If one looks at the map with the magnifying glass, then the map becomes a model of just one city. In the same way, at a low LoA the criteria of trustworthiness and benefits of trusting are specified, and so the definition at this level fits only a specific occurrence of trust or e-trust. At a high LoA the definition becomes:

> **Definition**: Assume a set of first order-relations functional to the achievement of a goal and that at least two agents (AAs or HAs) are involved in the relations, such that one of them (the trustor) has to achieve the given goal and the other (the trustee) is able to perform some actions in order to achieve that goal. If the trustor chooses to achieve its goal by the action performed by the trustee, and if the trustor considers the trustee a trustworthy agent, then the relation has the property of being advantageous for the trustor. Such a property is a second-order property that affects the first-order relations taking place between AAs, and is called trust.

This definition nowhere specifies criteria for the assessment of trustworthiness or the benefits of trust or e-trust in absolute terms. This general definition is the right definition of trust.

6. Conclusion

At the beginning of this paper I mentioned three aspects of the occurrence of e-trust that need to be taken in account by any analysis addressing this phenomenon: (i) the reasons for which an AA decides to trust, (ii) the effects of e-trust on the behaviour of the trustor, and (iii) the role of e-trust in the development of social interactions in distributed systems. The analysis of e-trust presented here contributes to explain all of these aspects.

In section 4 it has been shown that e-trust determines some advantages for the trustor and the achievement of its goal. The advantage is assessed considering the minimisation of the resources, i.e. time and energy, that the trustor spends in the process of accomplish its goal due to the presence of e-trust. Such advantage should be considered the reason for which a rational AA decides to trust another AA and take the (low) risk of being betrayed by the trustee. The choice to trust is simply the most convenient one, when the potential advantage determined by this choice is high and the potential risk of being deceived by the trustee is low.

The facilitating effect of e-trust also contributes to clarify the effects of e-trust on the trustor's behaviour. Once the trustor has chosen the trustee, then it can delegate the accomplishment of a given task to the trustee without having to supervise the other AA's performance. Delegation and absence of supervision should be considered, according to the analysis presented in this paper, the effects of e-trust on the trustor's behaviour.

Finally, if one considers the advantages determined by the occurrence of e-trust from a system-wise perspective then the role of e-trust in the development of social interactions in distributed systems becomes clearer. In section 1, I presented trust as a fundamental feature of social life, e-trust should be considered a phenomenon of the same importance for the development of social interactions in digital contexts. Such a fundamental role is a consequence of the benefits that e-trust confers on the trustor, as it makes advantageous for the agent who decides to trust to get involved in some relationships or in a net of relationships like a social system.

Going back to Hobbes, the advantages arising from *mutual trust* amongst the agents of a group are the reasons for choosing to end the status of *bellum omnium contra omnes* and establish a society. When e-trust occurs one can outsource part of one's duties to other agents and focuses on what one values more. e-Trust plays the same role in digital contexts, it promotes the delegation of tasks among the AAs of a system, in doing so it effectively reduces the complexity of the system (Luhmann 1979) and favours the development of social interactions.

References

Castelfranchi, C. and R. Falcone (1998). Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification. Third International Conference on Multi-Agent Systems. (ICMAS'98), Paris, France, IEEE Computer Society.

Corritore, C. L., B. Kracher, et al. (2003). "On-line trust: concepts, evolving themes, a model." International Journal of Human-Computer Studies **58**(6): 737-758

Floridi, L. (2008). "The Method of Levels of Abstraction." Minds and Machines **18**(3): 303-329.

Floridi, L. and J. Sanders (2004). "On the Morality of Artificial Agents." Minds and Machines **14**(3): 349-379.

Gambetta, D. (1998). Can We Trust Trust? Trust: Making and Breaking Cooperative Relations. D. Gambetta. Oxford, Basil Blackwell**:** 213–238.

Guttman, R., A. Moukas, et al. (1998). "Agent-Mediated Electronic Commerce: A Survey." Knowledge Engineering Review **13**(3): 147-159.

Lagenspetz, O. (1992). "Legitimacy and Trust." Philosophical Investigations **15**(1): 1-21.

Luhmann, N. (1979). Trust and Power. Chichester, Wiley.

Nissenbaum, H. (2001). "Securing Trust Online: Wisdom or Oxymoron." Boston University Law Review **81**(3): 635-664.

Nwana, H., J. Rosenschein, et al. (1998). Agent-Mediated Electronic Commerce: Issues, Challenges and some Viewpoints. Autonomous Agents 98, ACM Press.

Papadopoulou, P. (2007). "Applying virtual reality for trust-building e-commerce environments " Virtual Reality **11**(2-3): 107-127.

Seamons, K. E., M. Winslett, et al. (2003). Protecting Privacy during On-Line Trust Negotiation Privacy Enhancing Technologies. R. Dingledine and P. Syverson. Berlin / Heidelberg, Springer 249-253.

Taddeo, M. (2009). "Defining Trust and E-trust: Old Theories and New Problems,." nternational Journal of Technology and Human Interaction (IJTHI) **5**(2): 23-35.

Tuomela, M. and S. Hofmann (2003). "Simulating Rational Social Normative Trust, Predictive Trust, and Predictive Reliance between Agents." Ethics and Information Technology **5**(3): 163-176.

Vries, P. d. (2006). Social Presence as a Conduit to the Social Dimensions of Online Trust Persuasive Technology. W. IJsselsteijn, Y. d. Kort, C. Midden, B. Eggen and E. v. d. Hoven. Berlin / Heidelberg, Springer**:** 55-59.

Weckert, J. (2005). Trust in Cyberspace. The Impact of the Internet on Our Moral Lives. R. J. Cavalier. Albany, University of New York Press**:** 95-120.

Wooldridge, M. (2002). An introduction to multiagent systems. Chichester, J. Wiley.