# A cognitive trust and reputation model for the ART testbed

**Mario Gómez, Javier Carbó, Clara Benac Earle**

Department of Computer Science
University of Aberdeen, Scotland, UK
mgomez@csd.abdn.ac.uk
Departamento de Informática
Universidad Carlos III de Madrid
Av. Universidad 30, 28911, Leganés, Madrid (Spain)
{jcarbo,cbenac}@inf.uc3m.es

**Abstract**

Trust modelling is a challenging issue due to the dynamic nature of distributed systems and the unreliability of self-interested agents. In this context, the Agent Reputation and Trust (ART) Testbed has been used to compare trust models in two Spanish competitions in 2006 and 2007 and several international competitions. In this paper we describe the model we have presented to the Spanish competitions. We propose a trust model that uses the discrepancy between the information provided by other agents and its own experience in order to anticipate their actions, instead of using that discrepancy as a source of dishonesty and distrust. In addition we propose a cognitive model based on motivational attitudes to implement adaptive behaviors. Various implementations of this trust model have participated in the competitions; one was the winner of the first Spanish competition in 2006, the other won a combined game that included the participants in the two Spanish competitions.

## 1. Introduction

Trust is a universal concept that plays a very important role in social organizations as a mechanism for reducing the risk of social interactions and maximizing the utility obtained from such interactions. Therefore, modeling trust in open distributed systems such as agent systems becomes a critical issue since their offline and large-scale nature weaken the social control of direct interactions.

Sometimes there are objective and universal criteria to evaluate the quality of interactions (products/services provided by them). In such cases, trust can be inferred from certificates issued by third parties that verify such objective criteria. Unfortunately, there are many application domains in which the evaluation criteria is subjective (books, films, web pages, leisure activities, consulting services, technical assistance, etc.) and can be obtained only through local interactions. As a consequence, when a set of universal objective evaluation criteria is not available, agents have to relay on their own, subjective believes, to assess the trustworthiness of potential interaction partners.

One of the most effective ways to infer trust is through reputation-based mechanisms of social control [2]. Broadly explained, it consists of agents that form opinions about other agents

**ISSN: 1137-3601. ©AEPIA (http://erevista.aepia.org/)**

based on their past interactions, as well as from reports of third-party agents in order to improve the quality of decisions. Reputation is a concrete valuation that can be used to build trust in others. Usually, the group of people with good reputation (collaborators, colleagues and friends) that cooperates with a particular person to improve the quality of decisions forms an informal social network [20]. In this context, trust and reputation are strongly linked.

Due the many dimensions and application contexts, a breadth and diverse collection of trust and reputation models for multi-agent systems has been developed in recent years, without unified direction or benchmarks. Based on enthusiastic response from the agent trust community, the Agent Reputation and Trust (ART) testbed [7] (http://www.art-testbed.net/) initiative was launched in 2005, charged with the task of establishing a testbed for agent trust and reputation-related technologies. This testbed serves in two roles: (1) as a competition forum in which researchers can compare their technologies against objective metrics, and (2) as a suite of tools with flexible parameters, allowing researchers to perform customizable, easily-repeatable experiments.

In this paper we describe our particular trust model and how can it be tested using the ART testbed. Moreover, we comment on the results we have obtained in the 2006 and 2007 Spanish ART competitions.

The paper is organized as follows: section §2 discusses two trust models that have been proposed, followed by a description of our model and its mapping to the ART testbed in §3. Finally, §5 explains the results of our participation in the Spanish ART testbed competions in 2006 and 2007, and sums ups the contributions of our work.

## 2. Related work

Several trust models have been proposed; two of the most cited reputation models are SPORAS and HISTOS [22]. SPORAS is inspired in the foundations of the chess players evaluation system called ELOS. The key idea of this model is that trusted agents with very high reputation experience much smaller changes in reputation than agents with low reputation. SPORAS computes the reliability of agents' reputation using the standard deviation of such measure.

HISTOS is designed to complement SPORAS by including a way to deal with witness information (personal recommendations). HISTOS includes witness information as a source of reputation through a recursive computation of weighted means of ratings. It computes reputation of agent $i$ for agent $j$ from the knowledge of all the chain of reputation beliefs corresponding to each possible path connecting $i$ and $j$. In addition, HISTOS plans to limit the length of paths that are taken into account. To make a fair comparison with other proposals, that limit should be valued as 1, since most of the other views consider that agents communicate only their own beliefs, but not the beliefs of other sources that contributed to their own beliefs of reputation. Based on these principles, the reputation value of a given agent at iteration $i$, $R_i$, is obtained in SPORAS recursively from the previous one $R_{i-1}$ and from the subjective evaluation of the direct experience $DE_i$:

$$R_i = R_{i-1} + \frac{1}{\theta} \cdot \Phi(R_{i-1}) \cdot (DE_i - R_{i-1})$$

Let $\theta$ be the effective number of ratings taken into account in an evaluation ($\theta > 1$). The bigger the number of considered ratings, the smaller the change in reputation. Furthermore, $\Phi$ stands for a damping function that slows down the changes for very reputable users:

$$\Phi(R_{i-1}) = 1 - \frac{1}{1 + e^{\frac{-(R_{i-1}-D)}{\sigma}}}$$

where dominion $D$ is the maximum possible reputation value and $\sigma$ is chosen in a way that the resulting $\Phi$ would remain above 0.9 when reputations values were below $\frac{3}{4}$ of $D$.

Another well known reputation model is due to Singh and Yu. This trust model [21] uses Dempster-Shafer theory of evidence to aggregate recommendations from different witnesses. The main characteristic of this model is the relative importance of fails over success. It assumes that deceptions cause stronger impressions than satisfactions. It then applies different gradients to the curves of gaining/losing reputation in order to lose reputation easily, while it is hard to acquire it. The authors of this trust model define different equations to calculate reputation according to the sign (positive/negative) of the received direct experience (satisfaction/deception) and the sign of the previous reputation corresponding to the given agent.

Instead of Dempster-Shafer theory, Sen's reputation model [19] uses learning to cope with recommendations from different witnesses. Unfortunately learning requires a high number of interactions and a relatively high number of witnesses to avoid colluding agents benefiting from reciprocative agents.

Another remarkable reference in the field is ReGreT [17]. The ReGreT model takes into account three ways of computing indirect reputation depending on the information source: system, neighborhood and witness reputations. Note that witness reputation is the one that corresponds to the concept of reputation that we are considering. ReGreT includes a measure of the social credibility of the agent and a measure of the credibility of the information in the computation of witness reputation. The former is computed from the social relations shared between agents. It is computed in a similar way to neighborhood reputation, using third party references about the recommender directly in the computation of how its recommendations are taken into account. In addition, the latter measure of credibility (information credibility) is computed from the difference between the recommendation and what the agent experienced by itself. The similarity is computed by matching this difference with a triangle fuzzy set centered on 0 (the value 0 stands for no difference at all). The information credibility is considered relevant and taken into account in the experiments of this present comparison. Both decisions are also, to a certain degree, supported by the authors of ReGreT, who also assume that the accuracy of previous pieces of information (witness) are much more reliable than the credibility based on social relations (neighborhood), and they reduce the use of neighborhood reputation to those situations were there is not enough information on witness reputation. The complete mathematical expression of both measures can be found in [16]. But the key idea of ReGreT is that it also considers the role that social relationships may play. It provides a degree of reliability for the reputation values, and it adapts them through the inclusion of a temporal dependent function in computations. The time dependent function $\rho$ gives higher relevance to direct experiences produced at times closer to current time. The reputation held by any part at a iteration $i$ is computed from a weighted mean of the corresponding last $\theta$ direct experiences:

$$R_i = \sum_{j=i-\theta}^{j=i} \rho(i,j) \cdot W_j$$

where $W_j$ is an assessment of direct experiences for iteration $i$, and $\rho(i,j)$ is a weight applied to every $W_j$ based on the age of the experiences (older experiences have lower impact), and is calculated using the function below:

$$\rho(i,j) = \frac{f(j,i)}{\sum_{k=i-\theta}^{k=i} f(k,i)}$$

where $i \geq j$. Both represent the time or number of iterations of a direct experience. For instance, a simple example of a time dependent function f is:

$$f(j,i) = \frac{j}{i}$$

ReGreT also computes reliability with the standard deviation of reputation values, computed from:

$$STD - DVT_i = 1 - \sum_{j=i-\theta}^{j=i} \rho(i,j) \cdot \mid W_j - R_i \mid$$

ReGreT, however, defines reliability as a convex combination of this deviation with a measure, $0 < NI < 1$, whether the number of impressions, $i$, obtained is enough or not. ReGreT establishes an intimacy level of interactions, $itm$, to represent a minimum threshold of experiences to obtain close relationships. More interactions will not increase reliability. The next function models the level of intimacy with a given agent:

$$if(i \in [0, itm]) \rightarrow NI = \sin(\frac{\pi}{2 \cdot itm} \cdot i),$$
$$Otherwise \rightarrow NI = 1$$

Another trust model that is especially relevant to our work is due to Abdul-Rahmman [1]. This model does not take into account direct experiences, instead it draws on reputation as the only source of information to build trust. While other models combine witness information with direct infomation to infer trust on a target agent, this model uses direct experience to evaluate and correct the information provided by witnesses about the target agent. In this model, reputation is represented as a discrete belief with four possible values: very trustworthy, trustworthy, untrustworthy and very untrustworthy. The main contribution of this model is the use of the discrepancy between direct experience and witness information to infer trust on a third agent: prior to combine the information about a given agent provided by a number of witnesses, this information is adjusted according to previous information coming from that witness about any other agent and the experienced outcomes that support/invalidate such information. In other words, this model adjusts witness information according to the *semantic closeness* between the witness and oneself (how different they "think"); giving more importance to the information provided by agents that are perceived as being akin in terms of their cognitive models or preferences.

FIRE [11] is a trust and reputation model that integrates information coming from four different sources: interaction trust, role-based trust, witness reputation and certified reputation. Interaction trust is built from the direct experience of an agent, in particular, the direct trust component of ReGreT is exploited in this model. Role-based trust is based on relationships between the agents, which is mostly domain-specific. Witness information is built from reports of witnesses about an agent's behavior. Certified reputation is a novel type of reputation introduced by the authors, which is built from third-party references provided by the agent itself. Certified reputation plays a similar role to what we call advertisements, since in both cases an agent $i$ that has just joined the environment can make some assessment of the trustworthiness of another agent $j$, based on the certified reputation or advertisements provided by the agent $j$ itself. The main limitation of the FIRE model in [11] is that all agents are assumed to be honest in exchanging information.

Another approach when agents are acting in uncertain environments, is to apply adaptive filters such as Alpha Beta, Kalman and IMM [14, 5]. These filters have been recognized as a reasoning paradigm for time-variable facts within the Artificial Intelligence community [15]. Making time-dependent predictions in noisy environments is not an easy task. They apply a temporal statistical model to the noisy observations perceived through a linear recursive algorithm that estimates a future state variable. Particularly, when they are applied to reputation modeling, the state variable would be the reputation, while observations would be the results from direct experiences.

From an Artificial Intelligence perspective, reputation models embedded in agents should involve a cognitive approach[13]: enriching the internal model for making cooperative and competitive decisions rather than enriching the exchanged reputation information. In contrast to socio-cognitive models, computational models involve a numerical decision making, made up of utility functions, probabilities, and evaluations of past interactions. The combination of both computational models intends to reproduce the reasoning mechanisms behind human decision-making. In this paper we present a trust modeling framework that combines both views, since it assumes the cognitive stance, but uses a numerical approach.

Other researchers have proposed a *socio-cognitive* view of trust [12, 6, 3]. Schillo's model [12] distinguishes between two types of motivations for trust: honesty and altruism. A more enriched model is from Castelfranchi and Falcone [6] who claim that some other beliefs in addition to reputation are essential to compute the amount of trust of a particular agent: its competence (ability to act as we wish), willingness (intention to cooperate), persistence (consistency along time) and motivation (our contribution to its goals). For the authors, these beliefs should be taken into consideration in determining how much trust is set on this agent. Brainov and Sandholm [3] highlight the relevance of modeling opponent's trust, because, if this outside trust was not taken into account, this would lead to an inefficient trade between agents involved. Thus both agents would be interested in showing the trustworthiness of the counterpart to allocate efficiently resources.

Another example of a socio-cognitive approach is the fuzzy reputation agent system (AFRAS) by Carbo et al. [4], which supports the fuzzy nature of the reputation concept itself. It uses fuzzy logic to represent reputation since this concept is built up with vague evaluations (they depend on personal and subjective criteria), uncertain recommendations (malicious agents, different points of view), and incomplete information (untraceability of every agent in open systems). Furthermore, reliability of fuzzy reputation is implicit in the shape of the corresponding fuzzy set. Additionally it also includes other beliefs that intend to represent an emotive characterization of agents including shyness, egoism, and susceptibility. It also includes a global belief and a global adaptation value of agent interaction, referred to as *remembrance*. This attribute determines the relevance given to the last direct interaction when updating trustworthiness. It represents the general confidence of the agent on its own beliefs. The more success is achieved in predicting the behavior of a particular agent, the more relevance is applied to the already asserted beliefs over future experiences with any agent (not only that particular agent).

# 3. The DEAR trust and reputation model

A general definition of trust embraces many dimensions, for instance: to have belief or confidence in the honesty, goodness, skill or safety of a person, organization or thing (Cambridge Advanced Learners Dictionary). Herein we adopt an approach centered on the skill dimension: trust as a confidence in the competence, utility or satisfaction expected from other agent concerning a particular context. More specifically, we approach trust as a belief that estimates the quality of service (QoS) expected from a particular agent, based on both direct experience using that service, and information obtained from other agents.

We think that a combination of the numerical and the cognitive approaches discussed in the previous section is more appropriate to reproduce the reasoning mechanisms behind human decision-making. So we have conceived a trust and reputation framework that combines both views, since it assumes a cognitive stance, but uses a numerical representation.

Our global model of trust is based on a seamingless combination of elementary trust beliefs based on different sources of information. This model is called DEAR, that stands for Direct Experience, Advertisements and Recommendations based trust, which are the three sources of information we use to build trust.

## 3.1. Components of trust

Typically, a trust model considers two main sources of information: direct experience, which results in what is usually referred to as direct trust or interaction trust; and indirect experience, which is often referred to as witness-information, “word of mouth” or reputation. In our model we keep this distinction between direct and indirect experience but introduce a further distinction between the information provided by third party agents about other agents, what we call *recommendations*, and the information provided by an agent about itself, what we call *advertisements*. All in all, our trust model has three main components, namely: Direct Trust (DT), Advertisements-based Trust (AT), and Recommendations-based Trust (RT).

**Direct Trust($DT$):** assesses the Quality of Service (QoS) expected from an agent based on the past (direct) experience with that agent.

**Advertisements-based Trust ($AT$):** assesses the QoS expected from an agent based on the advertisements received from that agent, and the discrepancy between experience and advertisements observed in the past.

**Recommendations-based Trust ($RT$):** assesses the QoS expected from an agent based on the recommendations received from others about that agent, and the discrepancy between experience and recommendations observed in the past.

Some formal definitions follow:

**Direct Trust($DT_j^{\Sigma T}$):** assesses the QoS provided by agent $a_j$ until time step $T$ inclusive, based on direct experience.

$$DT_j^{\Sigma T} = \frac{\sum_{t=0}^{T} \varphi(t,T) pDT_j^t}{\sum_{t=0}^{T} \varphi(t,T)}$$

where $pDT_j^t : \mathbb{R} \to [0,1]$ is the partial DT obtained for $a_j$ in time step $t$, and $\varphi(T,t) : \mathbb{N} \to [0,1]$ is a forgetting function used to weight each partial belief according to its age (number of time steps since a belief was obtained, T-t). More comments on the forgetting functions can be found in Section 4.

**Direct Trust Confidence($DTC_j^{\Sigma T}$):** assesses the reliance of DT as an estimator of the QoS provided by agent $a_j$.

$$DTC_j^{\Sigma T} = ITM_j^{DT} \otimes (1 - \upsilon_{dt}(pDT_j^t))$$

where $ITM_j^{DT} \in [0,1]$ is the intimacy for $DT$ [17], a growing function in [0,1] over the number of $pDT$s used to compute $DT$, $\upsilon_{dt} \in [0,1]$ is a measure of the variability of $pDT_j^t$, and $\otimes$ is a *T-norm* operator.

**AT-Discrepancy $\Delta AT_j^{\Sigma T}$:** measures the discrepancy between the past advertisements made by agent $a_j$ and the experiences obtained.

$$\Delta AT_j^{\Sigma T} = \frac{\sum_{t=0}^{T} \varphi(t,T)(pDT_j^t - pAT_j^t)}{\sum_{t=0}^{T} \varphi(t,T)}$$

where $pAT_j^t : \mathbb{R} \to [0,1]$ is the Partial AT for agent $a_j$, and time step $t$, and $\varphi(t,T)$ is a time forgetting function.

Note that $\Delta AT_j^{\Sigma T} \in [-1,1]$, since $pDT_j^t, pAT_j^t, \varphi(t,T) \in [0,1]$ by definition. Positive values of $\Delta AT_j^{\Sigma T}$ means that the experiences obtained from agent $a_j$ were better than advertised, negative values have the opposite meaning, and zero means that the experiences matched perfectly with the advertisements.

**Advertisements-based Trust($AT_j^{T+1}$):** assesses the QoS expected from agent $a_j$ in the next time step $(T+1)$, based on the advertisements from $a_j$.

$$AT_j^{T+1} =$$

$$\left\{\begin{array}{ccc} 1 & if & pAT_j^{T+1} + \Delta AT_j^{\Sigma T} \geq 1 \\ 0 & if & pAT_j^{T+1} + \Delta AT_j^{\Sigma T} \leq 0 \\ pAT_j^{T+1} + \Delta AT_j^{\Sigma T} & if & 0 < pAT_j^{T+1} + \Delta AT_j^{\Sigma T} < 1 \end{array}\right\}$$

where $pAT_j^{T+1} : \mathbb{R} \rightarrow [0,1]$ is the most recent advertisement from $a_j$, and $\Delta AT_j^{\Sigma T}$ (AT-Discrepancy) is the discrepancy between advertisements and experiences obtained in the past (until time step $T$ inclusive).

**AT Confidence($ATC_j^{T+1}$):** assesses the degree of reliance on the AT as an estimation of the QoS to be obtained from agent $a_j$ in the next time step.

$$ATC_j^{T+1} = ITM_j^{AT} \otimes (1 - \upsilon_{at}(\Delta AT_j^t))$$

where $ITM_j^{AT}$ is the intimacy for $AT$, $\Delta AT_j^t = pDT_j^t - pAT_j^t$ is the partial discrepancy observed between $AT$ and $DT$ in time step $t$, $\upsilon_{at} \in [0,1]$ is a measure of the variability of $\Delta AT$, and $\otimes$ is a T-norm operator.

As we have done for DT and AT, we define both partial and historic Recommendations-based Trust (RT). However, RT must handle the fact that there are potentially many providers of information (recommenders) about any other agent. As a result, we have to distinguish between the trust component due to the recommendations provided by a single agent and the trust component due to the recommendations provided by several agents, what we call *combined recommendation.*

Finally, we have defined a global measure of trust that aggregates the components of trust into a single, global belief that we call Global Trust.

**Global Trust($GT_j^{T+1}$):** assesses the QoS expected from agent $a_j$ during the next time step, using all the sources of information.

$$GT_j^{T+1} =$$

$$\frac{DT_j^{\Sigma T} DTC_j^{\Sigma T} + AT_j^{T+1} ATC_j^{T+1} + cRT_j^{T+1} cRTC_j^{T+1}}{DTC_j^{\Sigma T} + ATC_j^{T+1} + cRTC_j^{T+1}}$$

where $DT_j^{\Sigma T}$ is the DT for agent $a_j$; $AT_j^{T+1}$ is the Anticipatory AT; $cRT_j^{T+1}$ is the Combined Recommendations-based Trust, and $DTC_j^{\Sigma T}$, $ATC_j^{T+1}$, $RTC_j^{T+1}$ are the confidences associated to $DT$, $AT$ and $cRT$ respectively.

**Global Trust Confidence($GTC_j^{T+1}$):** assesses the reliance on the Global Trust $GT_j$ as an estimation of the QoS to be obtained in the next time step.

$$GTC_j^{T+1} = DTC_j^{\Sigma T} \oplus ATC_j^{T+1} \oplus cRTC_j^{T+1} \quad (1)$$

where $\oplus$ is a *T-conorm* operator.

Essentially, the components of trust introduced above capture the skill dimension (QoS). To handle uncertainty and ignorance we introduce a measure of the reliability of a trust belief, what we call **confidence**. Actually, this belief breaks down into two components, namely: **intimacy**, and **predictability**. Intimacy is a measure of confidence based on the number of data (or interactions) used to calculate a belief, while predictability is a measure of confidence based on the dispersion or variability of the data. In our model, all the components of trust have attached a measure of confidence made up of intimacy and predictability. In addition, we propose the use of t-norms for combining intimacy and predictability into a single confidence value, and t-conorms for calculating the confidence coming from several sources of information (this is used to obtain the confidence on GT). The mathematical definition of all these beliefs and their corresponding confidence functions can be found in [10].

A key element of our model is how we handle discrepancy between information and direct experiences. In existing frameworks this discrepancy is used as a source of dishonesty: if an agent $i$ says that service $s$ has a quality of service $q$ and agent $j$ has experienced a quality of service $r$, then $q-r$ is assumed to represent a degree of dishonesty, a source of distrust. Unlikely, we propose a trust model which does not assume a concrete cognitive model for other agents an agent may interact with, but uses the discrepancy between the information provided by other agents and its own experience in order to anticipate their actions: $q-r$ is not used as a source of dishonesty and distrust, instead, it is used to estimate the quality of service to be obtained in the future. The result is an anticipatory trust modelling framework that allows agents to adapt swiftly to changes in the environment for its own benefit.

# 4. Trust dynamics

In [10] we have described the components of our trust model basically from an static viewpoint. Although the dynamic dimension is partially embraced in the notions of forgetting function and intimacy, the strongly dynamic nature of real open systems advocates for a more complex model to deal with and react to changes in the environment. Herein, we first review the dynamics in our previous model, and then describe a dynamic extension of the intimacy to deal with situations involving scarce or expensive information.

Due to the dynamic nature of open environments, it is common sense to assume that old beliefs, based on old perceptions, are less reliable that those beliefs based on recent perceptions. In consequence, several trust models, including ours, use a *forgetting function* to combine partial beliefs obtained at different times by weighting each belief according to its age. A concrete example of a forgetting function follows:.

$$\varphi(t,T) = \begin{Bmatrix} 0 & T-t \geq \phi \\ \cos(\frac{\pi}{2\phi}\cdot(T-t)) & 0 < T-t \end{Bmatrix}$$

where $t$ and $T$ are respectively a time in the past and the current time, and $\phi$ is a parameter that establishes the maximum age for an experience to be relevant (to influence current beliefs).

Nevertheless, the use of a forgetting function to weight each primitive component of a belief according to its age is not enough to account for a rapidly changing environment, actually, it is intended to aggregate the elementary components of a belief (single perceptions), given more importance to the more recent ones, but, what happens when there is no new information? In our previous model of trust there is an implicit assumption: that an agent has information and experience to update its beliefs each time step. However, the former assumption is quite simplistic and idealistic as to be applicable to the real world; on the contrary, in most real situations information is scarce or expensive, especially valuable information. Therefore, in general, we cannot assume that an agent will obtain information and experience continuously and exhaustively, instead we should devise a model to handle scarce information.

We have stated that each trust belief has associated a confidence value with two components: intimacy and predictability. Intimacy, as defined in [18, 8], encompasses dynamics in a very simplistic way: the intimacy is an ever growing function based on the number of data or interactions, until a maximum value is reached, and then intimacy becomes a constant. However, we claim that in order to deal with dynamic environments, intimacy should also decrease when there is no new information, to reflect the fact that an agent's knowledge becomes obsolete if it is not updated periodically.

To give a formal definition of the intimacy we need first to define a function that indicates whether a new perception has been obtained for a specific component of trust. Let $\Phi$ be any of the basic components of trust, i.e. $\Phi \in \{DT, AT, RT\}$; let $p\Phi$ be the partial value of $\Phi$ for a single time step, i.e. $p\Phi \in \{pDT, pAT, pRT\}$; and let $AV(p\Phi, T)$ be a boolean function that is true when $p\Phi$ is instantiated in time step $T$. Then, the intimacy for $\Phi$ in time step $T$, denoted by $ITM^{\Phi}(T)$, is defined recursively as a function of the intimacy at the previous time step $ITM^{\Phi}(T-1)$, and $AV(p\Phi,T)$

$$ITM^{T}(T) = \begin{Bmatrix} \Gamma^{\Phi} & if & T=0 \\ \Delta(ITM^{\Phi}(T-1)) & if & AV(p\Phi,T)=true \\ \Theta(ITM^{\Phi}(T-1)) & if & AV(p\Phi,T)=false \end{Bmatrix}$$

where $\Gamma^{\Phi} \in [0,1]$ plays the role of a prejudice, that is, a default value used to asses one's beliefs in the absence of other source of knowledge; $\Delta$ is a growing function $\Delta : [0,1] \rightarrow [0,1]$, and $\Theta$ is a decreasing function $\Theta : [0,1] \rightarrow [0,1]$.

In our experiments we have tried different functions to increment and decrement the intimacy, both additive and multiplicative, and in particular, for the experiments described later in this paper we have used functions of the following type:

$$\Delta(X) = X + \iota\ (1-X)$$
$$\Theta(X) = \delta\ X$$

where $\iota$ and $\delta$ are coefficients in $[0,1]$ (in addition we ensure that $ITM^{\Phi} \in [\Gamma^{\Phi},1]$).

Figure 1 depicts the kind of functions we are using; the functions at the left of the figure are examples of intimacy increment functions with different $\iota$ values (ranging from 0.2 to 0.5) , while the functions at the right are intimacy decrement functions with different $\delta$ values (ranging from 0.6 to 0.9). As new information is being received each turn, intimacy grows, tending to 1, and when no new information is received, the intimacy decreases. In general, we think that the increment pace for the intimacy should be considerably faster than the decrement pace. The idea here is that the occurrence of new information has a much stronger impact on the intimacy than the absence of information; which is suggested by the way human memory works. Anyway, we are not stating here that some specific functions are better than others to model the dynamics of the intimacy; instead, we advocate experimenting with different functions, as far as the dynamics of the intimacy follow this growing/falling mechanism based on the presence/absence of new information.

## 4.1. Motivational attitudes

While the former model represents a rational assessment of other agents' skills, in order to take intelligent decisions about whom to interact with and when, humans take into account also their particular needs and other motivational attitudes. Therefore, an agent decision-making may benefit from a human-like characterization of these cognitive dimensions.

Instead of adopting the binary-logic approach that is usual in the agent community –that of logic beliefs, desires and intentions–, we adopt a continuous approach to representing motivational attitudes, using variables in $[0, 1]$. In particular, we propose a model comprising five basic attitudes as the main forces driving the decision making of an agent, namely: *expected utility*, *necessity*, *knowledge/ignorance*, *satisfaction*, and *curiosity*

**Necessity:** assessment of the potential benefit that an agent could obtain by requesting a service from other agents; it is a measure of the lacking of skill, that is to say, the degree to which an agent is not utterly competent.

**Knowledge/Ignorance:** belief about the degree of knowledge (or absence of it) an agent has about a society of agents.

**Satisfaction/Dissatisfaction:** expresses the degree to which an agent is happy with the utility it expects to obtain from other agents in relation to its estimated necessity.

**Curiosity:** represents the attitude of an agent towards learning about other agents; it is defined as a function of the knowledge and the satisfaction, and more specifically, the curiosity is directly proportional to the ignorance and the dissatisfaction.

These attitudinal beliefs are defined below, but first we need to introduce another belief, the *expected utility*, which assesses the relative utility an agent expects from another agent compared with itself.

**Expected utility** ($U_j$) is the difference between the QoS an agent $a_i$ expects from another agent $a_j$ and the QoS that $a_i$ expects from itself.

$$U_j = (GT_j - myGT) \cdot A_j$$

where $GT_j$ is the Global Trust for agent $a_j$ , $myGT$ denotes the self-assessment of an agent about the QoS it is able to produce ($myGT \equiv GT_i$), and $A_j$ is the ratio of service requests accepted by $a_j$ (requests accepted/total requests).

The expected utility has attached a measure of confidence ($cU_j$) that is precisely the confidence value attached to $GT_j$, since the confidence on $myGT$ is assumed to be maximum (i.e. the null element for the conjunction: $cGT_j \otimes cGT_i = cGT_j$).

**Necessity** ($N$) is the difference between the maximum QoS potentially possible (a priori, without knowledge of other agents) and the QoS that an agent believes it is able to provide by itself.

$$N = 1 - myGT$$

**Knowledge** ($K$) is the proportion of intimacy an agent has compared with the maximum it is able to obtain.

$$K = \frac{\sum_{j=1}^{n} ITM_j^{GT}}{n-1}$$

where $ITM_j^{GT}$ is the intimacy of $GT_j$, and $n$ is the maximum number of agents that an agent is able or willing to know (with small agent societies $n$ could be the total number of agents, but with many agents $n$ must be limited). Complementarily, ignorance can be defined as $1 - K$

**Satisfaction** ($S$) is the complement of the difference between the lack of skill and the expected utility to be obtained from others.

$$S = 1 - (N - maxU) = (1 + maxU - N)$$

where $maxU$ represents the maximum utility an agent expects to obtain from others, and is defined as the $U_j \times cU_j$ that verifies $\forall i \neq j \mid U_i \times cU_i < U_j \times cU_j$. Complementarily, dissatisfaction can be defined as $1 - S$

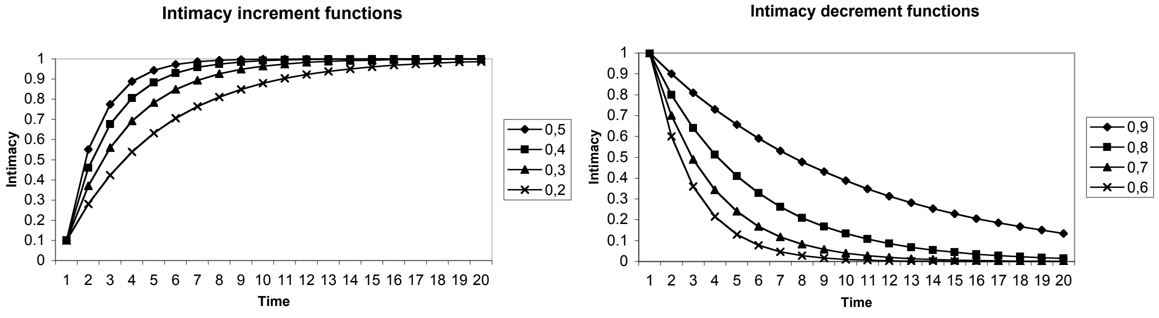

Figura 1: Example of intimacy increment and decrement functions

**Curiosity** ($C$) is the conjunction of ignorance and dissatisfaction.

$$C = (1 - S)(1 - K) = maxU(K - 1) + N(1 - K)$$

As stated above, the main purpose of these motivational attitudes is to drive the exploratory behavior of an agent, for instance, an agent with a high curiosity is highly motivated to know other agents, and thus, it would spend more resources performing epistemic actions.

Intuitively, an agent would have curiosity to know other agents either because it is ignorant about other agents, or because it is not satisfied with the QoS provided by the agents it already knows. An agent would not be curious when it thinks it does not need the services being provided by other agents (e.g. because it is already competent), or because it is quite satisfied with the service quality it is currently obtaining or expecting to obtain; in either case it is not worth spending resources to know new agents.

The use of curiosity and the other motivational attitudes may be used not only to drive the exploratory behavior of an agent, but also to decide on pragmatic actions. For example, necessity or satisfaction may be used to take decisions on the amount of delegation to conduct. An example of such an strategy is included in the empirical evaluation.

However, the use of curiosity is not the only approach to develop an exploratory behavior. The next section describes the empirical results obtained when comparing a curiosity-based strategy against several alternative strategies to explore the environment, using the ART testbed as the experimental platform.

## 4.2. Mapping to the ART testbed

The ART testbed is a simulator of the *art appraisals* domain whose goal is twofold: to serve as a competition forum in which researchers can compare their technologies against objective metrics, and as an experimental tool, with flexible parameters, allowing researchers to perform customizable, easily-repeatable experiments. In the art appraisal domain, agents act as painting appraisers with varying levels of expertise in different artistic eras (e.g. classical, impressionist, postmodern). Clients request appraisals for paintings from different eras. Appraisers can use both their own opinions and opinions purchased from other agents, so as to make more accurate appraisals. Appraisers estimate the accuracy of the opinions they send by the cost they choose to invest in generating an opinion, but they may lie about the estimated accuracy of their opinions. Appraisers receive more clients, and thus more profit, for producing more accurate appraisals. Appraisers may also purchase reputation information from other agents. The decisions about which opinion providers and reputation providers to trust strongly impact the accuracy of their final appraisals. In competition mode, the winning agent is selected as the appraiser with the highest bank account balance, which depends on the ability of an agent to (1) estimate the value of its paintings most accurately and (2) purchase more valuable information.

We map our trust model to the ART Testbed domain as follows:

- Direct Trust is mapped to the average error observed in the opinions an agent provides about the real value of a painting

$$\epsilon = \frac{|o-t|}{t}$$

where $o$ is the value of a painting according to an agent's opinion, and $t$ is the true value of that painting. The absolute value of the difference is used because we are measuring the magnitude of the error, not the direction. In addition, we invert the measure of error and enclose it in [0,1] to become a measure of trust: small errors implying a high $DT$, while large errors imply a low $DT$. We are using the following function to calculate the partial $DT$ for a single time step $T$ ($pDT^T$):

$$pDT^T = \left\{ \begin{array}{cc} 1 & \overline{\epsilon} \leq LB_\epsilon \\ 0 & \overline{\epsilon} \geq UB_\epsilon \\ \frac{UB_\epsilon - \overline{\epsilon}}{UB_\epsilon - LB_\epsilon} & 0{,}1 < \overline{\epsilon} < 1 \end{array} \right\}$$

where $LB$ and $UB$ are lower and upper bounds based on reasonable values for opinion errors. Values falling out of these limits are rated either as the best case (when $\overline{\epsilon} < LB$) or as the worst case (when $\overline{\epsilon} > UB$). The lower bound is actually posed by the simulation, and is 0.1, which is the minimum standard deviation of the errors and agent can have. There is no maximum standard deviation because the component $\frac{\alpha}{C_g}$ can be extremely big given that $C_g$ (the price invested to generate an opinion) can be arbitrarily small. However a reasonable upper limit for $\frac{\alpha}{C_g}$ can be established depending on the expected domain of $C_g$ in the context of an specific combination of simulation parameters; for instance, we can assume than an agent will never spend more money generating opinions than the fee $f$ paid by clients per appraisal. In particular, in our experiments we have been using 0.1 or 0.2 as reasonable upper limits for $\frac{\alpha}{C_g}$, which implies $UB = 1{,}1$ or $UB = 1{,}2$ respectively.

- Advertisements are mapped to `certainties`. Certainties are values in [0,1] provided by agents when their opinions are requested. These values are supposed to represent the accuracy of the opinions to be provided by an agent in a given era. We aggregate all the certainties provided by an agent for an era during a single timestep using the mean.
- Recommendations are mapped to `reputations`. There is a direct mapping between a single reputation and what we call partial RT ($pRT$), since a reputation is a value in [0,1] that represents the assessment on the opinions accuracy of other agents for a specific era.
- The Global Trust × Global Trust Confidence ($GT \times GTC$) is mapped to the weight an agent has to provide every timestep for every agent including itself. This value represents the final trust assessment used together with the motivational attitudes to decide from which agents to request opinions and reputations every time step.

# 5. Competition results and conclusions

In this paper we have presented a short description of the model that we have developed and tested using the ART testbed. The underlying mathematical models and more detailed discussions of the benefits and drawbacks of our approach can be found in [10, 9], together with an analysis of some experimental results.

Here we present the results obtained by our agent in the Spanish competitions as additional evidence of the advantages of our approach: a first implementation of our model won the First Spanish ART Competition[1] (agent named Basic in Figure 2b), and won the combined game conducted during the Second Spanish Competition[2], which involved all the participants in any of the two Spanish competitions (agent named Guzman in Figure 2a). Note that our agent in the First Competition didn't use reputation information, for it was not worth given that only a handful of agents partook in any single game. This situation contrasts with the combined game conducted during the Second Competition, which involved more than 20 participants, and was won by an implementation of the same agent model using reputation (Guzman). This result may indicate the convenience of using reputation information in large agent societies, for in that case knowing every other agent through direct experience is unfeasible or too expensive. However, we should conduct more experiments in order to obtain statistically relevant information to support such a

[1] `http://www.giaa.inf.uc3m.es/ART_Testbed_SPAIN/los_resultados.htm`

[2] `http://megatron.iiia.csic.es/spanishART/`

claim.

Thanks to the experiments discussed in [8] we have proven that our model performs better than other state-of-art models when facing dynamic environments, which is not the case of the ART testbed version used for the Spanish competitions. However, the use of the attitudinal model has also proven very useful as a mechanism of developing smart exploratory behaviors that turn to result in a more efficient behavior (better ratio between incomes and expenses). This is probably the main factor explaining our success in the competitions.

# Acknowledgements

Partly funded by projects COLOCAME CICYT TSI2005-07344-C02-02, AUTOPIA (IMSERSO 31/06) and SINPROB (FOMENTO T10/2006).

# Referencias

[1] Alfarez Abdul-Rahman and Stephen Hailes. Supporting trust in virtual communities. In *HICSS '00: Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 6*, page 6007, Washington, DC, USA, 2000. IEEE Computer Society.

[2] S. Ba, A.B. Whinston, and H. Zhang. Building trust in online auction markets through an economic incentive mechanism. *Decision Support System*, 35:273–286, 2003.

[3] S. Braynov and T. Sandholm. Trust revelation in multiagent interaction. In *Workshop on The Philosophy and Design of Socially Adept Technologies, Minneapolis*, pages 57–60, 2002.

[4] J. Carbo, J.M. Molina, and J. Davila. Trust management through fuzzy reputation. *International Journal of Cooperative Information Systems*, 12(1):135–155, March 2003.

[5] Javier Carbo, Jesus Garcia, and Jose M. Molina. Convergence of agent reputation with alpha-beta filtering vs. a fuzzy system. In *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce Vol-1*, pages 327–331, Washington, DC, USA, 2005. IEEE Computer Society.

[6] C. Castellfranchi and R. Falcone. Principles of trust for multiagent systems: Cognitive anatomy, social importance and quantification. In *Third International Conference on Multi-Agent Systems*, pages 72–79, 1998.

[7] K. Fullam, T. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. Rosenschein, L. Vercouter, and M. Voss. A specification of the agent reputation and trust (art) testbed: Experimentation and competition for trust in agent societies. In *The Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2005)*, pages 512–518, 2005.

[8] Mario Gomez, Javier Carbo, and Clara Benac. An anticipatory trust model for open distributed systems. volume 4520 of *Lecture Notes in Computer Science*, pages 307–324. Springer, 2007.

[9] Mario Gomez, Javier Carbo, and Clara Benac. Trust dynamics, motivational attitudes and epistemic actions: a curiosity-driven exploratory behavior. In *Proceedings of the Tenth Workshop on Trust and Reputation, held at the Autonomous Agents and Multi-Agent Systems (AAMAS-2007)*, pages 68–79, Honolulu, Hawaii, USA, 2007.

[10] Mario Gómez, Javier Carbó, and Clara Benac Earle. Honesty and trust revisited: the advantages of being neutral about other's cognitive models. *Autonomous Agents and Multi-Agent Systems*, 15(3):313–335, 2007.

[11] Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[12] P. Funk M. Schillo and M. Rovatsos. Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence, Special Issue on Trust, Deception and Fraud in Agent Societies*, 14(8):825–849, 2000.

[13] S. Marsh. Trust in distributed artificial intelligence. In Castelfranchi and Werner, editors, *Lecture Notes in Artificial Intelligence 830*, pages 94–112. Springer Verlag, 1994.

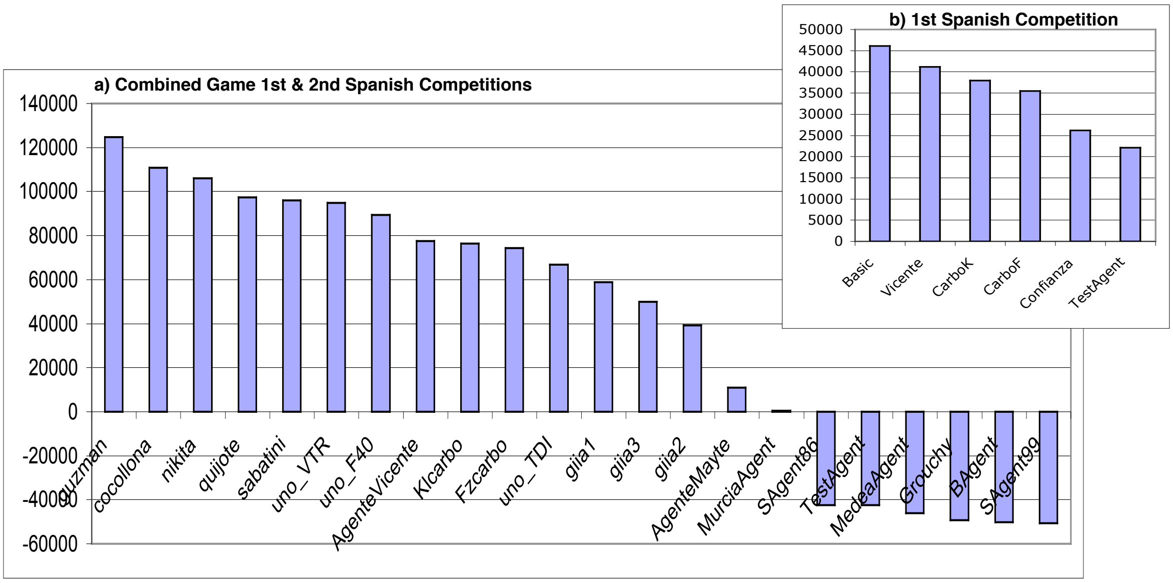

Figura 2: Summary of competition results

[14] Javier Carbo Rubiera, Jesus Garcia, and Jose M. Molina. Subjective trust inferred by kalman filtering vs. a fuzzy reputation. In *Conceptual Modeling for Advanced Application Domains*, volume 3289 of *Lecture Notes in Computer Science*, pages 496–505. Springer, 2004.

[15] S.J. Russell and P.Ñorvig. *Artificial intelligence: a modern approach.* Prentice Hall Pearson Education International, 2003.

[16] J. Sabater. *Trust and Reputation for agent societies. Ph.D. Thesis.* Consejo Superior de Investigaciones Cientificas, Bellaterra, Spain, 2003.

[17] J. Sabater and C. Sierra. Regret: a reputation model for gregarious societies. In *Fourth Workshop on Deception, Fraud and Trust in Agent Societies*, pages 61–69, Montreal, Canada, 2001.

[18] J. Sabater and C. Sierra. Reputation and social network analysis in multiagent systems. In *First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS2002)*, pages 475–482, Bologna, Italy, 2002.

[19] S. Sen, A. Biswas, and S. Debnath. Believing others: pros and cons. In *Proceedings of the 4th International Conference on MulitAgent Systems*, pages 279–285, Boston, MA, July 2000.

[20] S. Wasserman and J. Galaskiewicz. *Advances in Social Network Analysis.* Sage Publications, Thousand Oaks, U.S., 1994.

[21] B. Yu and M.P. Singh. A social mechanism for reputation management in electronic communities. *Lecture Notes in Computer Science*, 1860:154–165, 2000.

[22] G. Zacharia and P. Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, 14:881–907, 2000.