The examination of an information-based approach to trust

Maaike Harbers[1], Rineke Verbrugge[2], Carles Sierra[3], and John Debenham[4]

[1] Institute of Information and Computing Sciences, Utrecht University, P.O.Box 80.089, 3508 TB Utrecht, The Netherlands
maaike@cs.uu.nl
[2] Institute of Artificial Intelligence, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands
rineke@ai.rug.nl
[3] IIIA-CSIC, Campus UAB, 08193 Cerdanyola, Catalonia, Spain
sierra@iiia.csic.es
[4] Faculty of Information Technology, University of Technology, Sydney, PO Box 123, Broadway, NSW 2007, Australia
debenham@it.uts.edu.au

Abstract. This article presents the results of experiments performed with agents based on an operalization of an information-theoretic model for trust. Experiments have been performed with the ART test-bed, a test domain for trust and reputation aiming to provide transparent and recognizable standards. An agent architecture based on information theory is described in the paper. According to a set of experimental results, information theory is shown to be appropriate for the modelling of trust in multi-agent systems.

1 Introduction

In negotiation, one tries to obtain a profitable outcome. But what is a profitable outcome: to pay little money for many goods of high quality? Although this seems to be a good deal, it might not always provide the most profitable outcome in the long run. If negotiation partners meet again in the future, it could be more rational to focus on the relationship with the other agents, to make them trust you and to build up a good reputation.

In computer science and especially in distributed artificial intelligence, many models of trust and reputation have been developed over the last years. This relatively young field of research is still rapidly growing and gaining popularity. The aim of trust and reputation models in multi-agent systems is to support decision making in uncertain situations. A computational model derives trust or reputation values from the agent's past interactions with its environment and possible extra information. These values influence the agent's decision-making process, in order to facilitate dealing with uncertain information.

Big differences can be found among current models of trust and reputation, which indicates the broadness of the research area. Several articles providing an overview of the field conclude that the research activity is not very coherent and needs to be

more unified [1–4]. In order to achieve that, test-beds and frameworks to evaluate and compare the models are needed.

Most present models of trust and reputation make use of game-theoretical concepts [1, 5]. The trust and reputation values in these models are the result of utility functions and numerical aggregation of past interactions. Some other approaches use a cognitive model of reference, in which trust and reputation are made up of underlying beliefs. Castelfranchi and Falcone [6] developed such a cognitive model of trust, based on beliefs about competence, dependence, disposition, willingness and persistence of others. Most existing models of trust and reputation do not differentiate between trust and reputation, and if they do, the relation between trust and reputation is often not explicit [1, 3]. The ReGreT system [7] is one of the few models of trust and reputation that does combine the two concepts. Applications of computational trust and reputation systems are mainly found in electronic markets. Several research reports have found that seller reputation has significant influences on on-line auction prices, especially for high-valued items [3]. An example is eBay, an online market place with a community of over 50 million registered users [2].

Sierra and Debenham [8] introduced an approach using information theory for the modeling of trust, which has been further developed in [9], [10]. The present article presents an examination of Sierra and Debenham's information-based approach to trust. Experiments have been performed with the ART test-bed [4], a test domain for trust and reputation. Section 2 introduces the trust model, section 3 describes the ART test-bed, and section 4 describes how the model has been translated into an agent able to participate in the ART test-bed. The remainder of the article gives an overview of the experiments (section 5) and the results (section 6), followed by a discussion (section 7). The article ends with conclusions and recommendations for further research (section 8).

2 The information-based model of trust

In Sierra and Debenham's information-based model, trust is defined as the measure of how uncertain the outcome of a contract is [8]. All possible outcomes are modelled and a probability is ascribed to each of them. More formally, agent α can negotiate with agent β and together they aim to strike a deal δ. In the expression $\delta = (a, b)$, a represents agent α's commitments and b represents β's commitments in deal δ. All agents have two languages, language C for communication and language L for internal representation. The language for communication consists of five illocutionary acts (Offer, Accept, Reject, Withdraw, Inform), which are actions that can succeed or fail. With an agent's internal language L, many different worlds can be constructed. A possible world represents, for example, a specific deal for a specific price with a specific agent.

To be able to make grounded decisions in a negotiation under conditions of uncertainty, the information-theoretic method denotes a probability distribution over all possible worlds. If an agent would not have any beliefs or knowledge, it would ascribe to all worlds the same probability to be the actual world. Often however, agents do have knowledge and beliefs which put constraints on the probability distribution. The agent's knowledge set K restricts *all worlds* to all *possible worlds*: that is, worlds that are consistent with its knowledge. Formally, a world v corresponds to a valuation function on

the positive ground literals in the language, and is an element of the set of all possible worlds V. Worlds inconsistent with the agent's knowledge are not considered.

An agent's set of beliefs B determines its opinion on the probability of possible worlds: according to its beliefs some worlds are more probable to be the actual world than others. In a probability distribution over all possible worlds, W, a probability p_i expresses the degree of belief an agent attaches to a world v_i to be the actual world. From a probability distribution over all possible worlds, the probability of a certain sentence or expression in language L can be derived. For example the probability $P(\textit{executed} \mid \textit{accepted})$ of whether a deal, once accepted, is going to be executed can be calculated. This derived sentence probability is considered with respect to a particular probability distribution over all possible worlds. The probability of a sentence σ is calculated by taking the sum of the probabilities of the possible worlds in which the sentence is true. For every possible sentence σ that can be constructed in language L the following holds: $P_{\{W|K\}}(\sigma) \equiv \Sigma_n\{p_n : \sigma \text{ is true in } v_n\}$. An agent has attached given *sentence probabilities* to every possible statement φ in its set of beliefs B.

A probability distribution over all possible worlds is consistent with the agent's beliefs if for all statements in the set of beliefs, the probabilities attached to the sentences are the same as the derived sentence probability. Expressed in a formula, for all beliefs φ in B the following holds: $B(\varphi) = P_{\{W|K\}}(\varphi)$. Thus, the agent's beliefs impose linear constraints on the probability distribution. To find the best probability distribution consistent with the knowledge and beliefs of the agent, *maximum entropy inference* (see [11]) uses the probability distribution that is maximally non-committal with respect to missing information. This distribution has maximum entropy and is consistent with the knowledge and beliefs. It is used for further processing when a decision has to be made.

When the agent obtains new beliefs, the probability distribution has to be updated. This happens according to the principle of *minimum relative entropy*. Given a prior probability distribution $\underline{q} = (q_i)_{i=1}^n$ and a set of constraints, the *principle of minimum relative entropy* chooses the posterior probability distribution $\underline{p} = (p_i)_{i=1}^n$ that has the least relative entropy with respect to $\underline{q}$, and that satisfies the constraints. In general, the relative entropy between probability distribution p and q is calculated as follows: $D_{RL}(p \parallel q) = \Sigma_{i=1}^n p_i \log_2 \frac{p_i}{q_i}$. The principle of minimum relative entropy is a generalization of the principle of maximum entropy. If the prior distribution $\underline{q}$ is uniform, the relative entropy of $\underline{p}$ with respect to $\underline{q}$ differs from the maximum entropy $H(\underline{p})$ only by a constant. So the principle of maximum entropy is equivalent to the principle of minimum relative entropy with a uniform prior distribution (see also [8]).

While an agent is interacting with other agents, it obtains new information. Sierra and Debenham [8] mention the following types of information from which the probability distribution can be updated:

- *Updating from decay and experience*. This type of updating takes place when the agent derives information from its direct experiences with other agents. It is taken into account that negotiating people or agents may forget about the behavior of a past negotiation partner.
- *Updating from preferences*. This updating is based on past utterances of a partner. If agent α prefers a deal with property Q_1 to a deal with Q_2, he will be more likely to accept deals with property Q_1 than with Q_2.

– *Updating from social information*. Social relationships, social roles and positions held by agents influence the probability of accepting a deal.

Once the probability distribution is constructed and up to date, it can be used to derive trust values. From an actual probability distribution, the trust of agent α in agent β at the current time, with respect to deal δ or in general, can be calculated. The trust calculation of α in β is based on the idea that the more the actual executions of a contract go in the direction of the agent α's preferences, the higher its level of trust. The relative entropy between the probability distribution of acceptance and the distribution of the observation of actual contract execution models this idea. For $T(\alpha,\beta,b)$, the trust of agent α in agent β with respect to the fulfillment of contract (a,b), the following holds:

$$T(\alpha,\beta,b) = 1 - \sum_{b' \in B(b)^+} P^t(b') \log \frac{P^t(b')}{P^t(b'|b)}$$

Here, $B(b)^+$ is the set of contract executions that agent α prefers to b. $T(\alpha,\beta)$, the trust of α in β in general, is the average over all possible situations. After making observations, updating the probability distribution and calculating the trust, the probability of the actual outcomes for a specific contract can be derived from the trust value and an agent can decide about the acceptance of a deal.

3 The ART Test-bed

Participants in the ART test-bed [4] act as appraisers who can be hired by clients to deliver appraisals about paintings, each for a fixed client fee. Initially, a fixed number of clients is evenly distributed among appraisers. When a session proceeds, appraisers whose final appraisals were most accurate are rewarded with a larger share of the client base. Each painting in the test-bed has a fixed value, unknown to the participating agents. All agents have varying levels of expertise in different artistic eras (e.g. classical, impressionist, post-modern), which are only known to the agents themselves and which will not change during a game. To produce more accurate appraisals, appraisers may sell and buy opinions from each other. If an appraiser accepts an opinion request, it has to decide about how much time it wants to invest in creating an opinion. The more time (thus money) it spends in studying a painting, the more accurate the opinion.

However, agents might (on purpose) provide bad opinions or not provide promised opinions at all. Then without spending time on creating an opinion, the seller receives payment. So to prevent paying money for a useless opinion, the test-bed agents have to learn which agents to trust. To facilitate this process, agents can buy information about other agents' reputations from each other. Here again agents do not always tell the truth or provide valuable information.

Appraisers produce final appraisals by using their own opinion and the opinions received from other appraisers. An agent's final appraisal is calculated by the simulation, to ensure that appraisers do not strategize for selecting opinions after receiving all purchased opinions. The final appraisal $p*$ is calculated as a weighted average of received opinions: $p* = \frac{\sum_i (w_i \cdot p_i)}{\sum_i w_i}$. In the formula, p_i is the opinion p received from provider i and w_i is the appraiser's weight for provider i: the better α trusts an agent i, the higher the weight w_i attached to that agent and the more importance will be given to its opinion.

Agent α determines its final appraisal by using all the opinions it received plus its own opinion. The true painting value t and the calculated final appraisal $p*$ are revealed by the simulation to the agent. The agent can use this information to revise its trust models of other participants.

4 An information-based test-bed agent

The implemented test-bed agent ascribes probabilities to the accuracy of the opinions other agents provide. The agent maintains a probability distribution for each era of expertise with respect to each agent. The different possible worlds in a probability distribution represent the possible grades of the opinions an agent might provide in a specific era. An opinion of high grade means that the appraised value of a painting is close to the real value of the painting. A low grade means that the agent provides very bad opinions in the corresponding era or that the agent does not provide opinions at all. The quality of an opinion actually is a continuous variable, but to fit the model all possible opinions are grouped into ten levels of quality. The act of promising but not sending an opinion is classified in the lowest quality level.

The probability distributions are updated during the course of a session each time the agent receives new information, which can be of three types:

- Updating from direct experiences;
- Updating from reputation information;
- Updating from the evaporation of beliefs (forgetting).

Updating from reputation information corresponds to *Updating from social information* in Sierra and Debenham's model [8]. The other two types of updating are derived from *Updating from decay and experience* in the model.

Updating from direct experiences takes place when the agent receives the true values of paintings. The value of a constraint is obtained by taking the relative error of an opinion: the real value of a painting and an agent's estimated value of a painting are compared to each other. *Updating from reputation information* takes place when the agent receives witness information. The value of a constraint is derived by taking the average of the reputation values in all messages received at a specific time from trusted agents about a specific agent and era. *Updating from forgetting* is performed each time when a probability distribution is updated either from direct experiences or from reputation information.

Direct experiences and reputation information are translated into the same type of constraints. Such a constraint is for example: agent α will provide opinions with a quality of at least 7 in era e with a certainty of 0.6. This constraint is put to the probability distribution of agent α and era e. After updating from this constraint, the probabilities of the worlds 7, 8, 9 and 10 should together be 0.6. Constraints are always of the type opinions of *at least* quality x.

The value of a constraint (the quality grade) derived from a direct experience is obtained by comparing the real value of a painting to an agent's estimated value according to the equation: $constraintValue = 10 \cdot (1 - \frac{|appraisedValue - trueValue|}{trueValue})$. The outcome represents the quality of the opinion and a new constraint can be added to the set of beliefs. If a value lower than one is found, a constraint with the value of one is added to the set of beliefs. Reputation information is translated into a constraint by taking the average

of the reputation values in all messages received at a specific time from trusted agents about a specific agent and era multiplied by ten: $constraintValue = 10 \cdot \Sigma_{r \in reps} \frac{r}{n_1}$, where r is a reputation value, *reps* is the set of useful reputation values and n_1 is the size of *reps*.

With a set of constraints and the principle of maximum entropy, an actual probability distribution can be calculated. Therefore one general constraint is derived from all the stored constraints for calculating the probability distribution. The general constraint is a weighted average of all the constraints stored so far, calculated according to the following equation: $generalconstraintValue = \frac{1}{n_2} \cdot \Sigma_{c \in C} \frac{1}{(c(t_{obtained}) - t_{current}) + 1} \cdot c(value)$, where constraint c is an element of the set C of stored constraints and n_2 the total amount of constraints. Each constraint c consists of the time it was obtained $c(t_{obtained})$ and a quality grade $c(value)$, calculated with one of the formulas *constraintValue* above. The outcome is rounded to get an integer value.

The constraints are weighted with a factor of one divided by their age plus one (to avoid fractions with a zero in the denominator). Forgetting is modelled by giving younger constraints more influence on the probability distribution than older constraints. In this calculation, constraints obtained from reputation information are weighted with a factor which determines their importance in relation to constraints obtained from direct information. A ratio of 0.3:1, respectively, was taken because reputation info is assumed to have less influence than info from direct experiences. With the principle of maximum entropy, a new and updated probability distribution can be found.

Finally, when all information available has been processed and the probability distributions are up to date, trust values can be derived from the probability distributions. There are two types of trust, the trust of a particular agent in a specific era and the trust of a particular agent in general. The trust value of an agent in a specific era is calculated from the probability distribution of the corresponding agent and era. In an *ideal probability distribution*, the probability of getting opinions of the highest quality is very high and the probability of getting opinions with qualities lower than that is very low. Now trust can be calculated by taking one minus the relative entropy between the ideal and the actual probability distribution, as follows: $trust(agent, era) = 1 - \Sigma_{i=1}^{n_3} (P_{actual}(i) \cdot \log \frac{P_{actual}(i)}{P_{ideal}(i)})$, where n_3 is the number of probabilities. The trust of an agent in general is calculated by taking the average of the trust values of that agent in all the eras. At each moment of the game, the agent can consult its model to determine the trust value of an agent in general or the trust value of an agent with respect to a specific era. These trust values guide the behavior of the agent.

At the beginning of a new session the agent trusts all agents, so the probability distributions are initialized with all derived trust values (for each agent in each era) at 1.0. During the game the model is updated with new constraints and trust values change. The general behavior of the information-based agent is honest and cooperative towards the agents it trusts. The agent buys relevant opinions and reputation messages from all agents it trusts (with trust value 0.5 or higher). The agent only accepts and invests in requests from trusted agents, and if the agent accepts a request it provides the best possible requested information. If the agent does not trust a requesting agent, it informs the other agent by sending a decline message. If a trusted agent requests for reputation information, the agent provides the trust value its model attaches to the subject agent.

If the agent trusts an agent requesting for opinions, it always highly invests in ordering opinions from the simulator for that agent. Finally, the agent uses the model for generating weights for calculating the final opinions. It weights each agent (including itself) according to the trust in that agent in that era.

5 Set-up of the experiments

To test the influences of the use of different types of information, four variations of an information-based agent have been made. The suffixes in the names of the agents indicate the information types they use for updating: *de* corresponds to direct experiences, *rep* to reputation information and *time* to forgetting.

- Agent *Info-de* only updates from direct experiences;
- Agent *Info-de-time* updates from direct experiences and from forgetting;
- Agent *Info-rep-time* updates from reputation information and forgetting;
- Agent *Info-de-rep-time* updates from all three types of information.

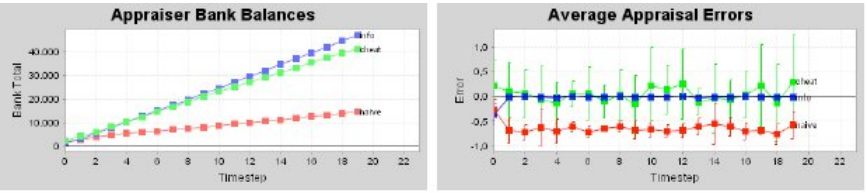
The performances of these agents in the ART test-bed are in the first place measured by their ability to make accurate appraisals, which is indicated by their client shares after the last game round. Besides, information about the agents' bank account balances will be presented. The use of each of the information types is expected to increase the average appraisal accuracy of an information-based test-bed agent. Moreover, the use of the combination of all three information types is expected to deliver the best results. In order to verify the correctness of these expectations, three test conditions have been designed and four extra agents have been implemented.

The **first condition** tests an agent's ability to distinguish between a cooperating and a non-cooperating agent. In this first part of the experiment, the agents *Info-de*, *Info-de-time* and *Info-de-rep-time* each participated in a game together with the test-agents *Cheat* and *Naive*. The test-agent *Cheat* never makes reputation or opinion requests itself, but when it receives requests it always promises to provide the requested reputation information or opinions. As its name suggests, the agent cheats on the other agents and it never sends any promised information. Its final appraisals are just based on its own expertise. The agent *Naive* bases its behavior on the idea that all agents it encounters are trustworthy and *Naive* keeps on trusting others during the whole course of a game. This agent always requests every other agent for reputation information and opinions, it accepts all requests from other agents and it highly invests in creating the requested opinions. Its final appraisals are based on its own expertise and on the (promised but sometimes not received) opinions of all other agents.

For the **second condition**, a third test-agent was developed to investigate other agents' ability to adapt to new situations. This agent *Changing* shows the same behavior as *Naive* during the first ten rounds of a game. Then it suddenly changes its strategy and from the eleventh game round till the end of the game it behaves exactly the same as the agent *Cheat*. The performances of the agents *Info-de* and *Info-de-time* in reaction to *Changing* have been examined.

The **third condition** was designed to examine the updating from reputation information. This type of updating is only of use if there are agents in the game that provide reputation information, so a reputation information providing agent *Providing* has been

Fig. 1. Bank account balances and average appraisal errors of agents *Info-de-time* (black), *Cheat* (light grey) and *Naive* (dark grey) in the first test conditions.



	Cheat		**Naive**		**Agent**	
	Bank	*Client*	*Bank*	*Client*	*Bank*	*Client*
info-de	45957	24.5	14361	8.8	40700	26.4
info-de-time	47975	25.9	13552	8.8	40262	25.0
info-de-rep-time	46097	24.7	14073	8.2	41461	26.7

Table 1. Averages for three information-based agents in conditions of type one.

implemented. The only difference with *Info-de-time* is that the *Providing* agent always accepts reputation requests and provides the wished reputation information, whereas the agent *Info-de-time* only provides reputation to agents it trusts. The agents *Info-de-time*, *Info-rep-time* and *Info-de-rep-time* each participated in a game with *Providing*, *Cheat* and *Naive*.

6 Results

In the first experiment, each of the agents *Info-de*, *Info-de-time* and *Info-de-rep-time* participated in a test-bed game together with the agents *Cheat* and *Naive*. The graphics in Figure 1 show an example of a session with the agents *Info-de-time*, *Cheat* and *Naive*. Left the development of the agents' bank account balance during the whole game is shown. All agents have increasing balances, but *Info-de-time* ends the game with the most and *Naive* with the least money. The right part of the figure shows the average appraisal errors of the agents in each round. The appraisals of *Naive* are obviously less accurate than those of the other two agents. This can be explained by *Naive*'s behavior to keep on trusting the cheating agent during the whole game. *Info-de-time* provides its least accurate appraisals the first game round; there it still has to learn that it cannot trust the agent *Cheat*. After that, its appraisals are the most accurate: the errors are close to the zero line and show the least deviation. This can be explained by *Info-de-time* using the expertise of two agents (itself and *Naive*), whereas *Cheat* only uses its own expertise.

Table 1 shows the averages of 30 sessions for the three information-based agents in condition one. In the tables, Client refers to the final number of clients of an agent at the end of a session and Bank means its final bank account balance. The first row shows the average final bank account balance and average final number of clients of respectively, *Cheat*, *Naive* and *Info-de*, for the sessions in which the three of them participated together in the game. The second row displays the results of the sessions with *Cheat*,

Naive and *Info-de-time*. Applying Student T-test (two-tailed, homoscedastic distribution) showed that with a significance level of 5% one can only conclude that *Info-de-rep-time* gathers a significantly bigger client share than *Info-de-time*. The differences in bank account balances between the different agents are not significant.

In the second condition *Info-de* and *Info-de-time* participate in a game with the agent *Changing*, which starts to cheat from the tenth round of the game. In contrast to *Info-de*, the agent *Info-de-time* does take forgetting into account. As time goes by, information gathered in the past becomes less and less important. The difference is clear: after a first big decrease in appraisal accuracy when the agent *Changing* starts cheating, *Info-de-time* learns from *Changing*'s new behavior and adjusts its trust values. Its past beliefs about a seemingly trustworthy agent *Changing* do not overrule the new information it gathers and it ends with higher scores. The averages of all the sessions with the agent *Changing* are presented in Table 2. Both client share and bank account balance of the two information-based agents are significantly different on a 5% level of significance according to the Student T-test. The results of the third condition, testing the update from reputation information, are shown in Table 3. A Student T-test demonstrates that all differences in client shares between the three tested agents are significant.

	Changing		**Agent**	
	Bank	*Client*	*Bank*	*Client*
info-de	44189	33.4	25817	6.6
info-de-time	36211	21.2	33864	18.8

Table 2. Averages for the agent *Changing*.

7 Discussion

It was expected that the experiments would show that each of the three types of updating would contribute to appraisal accuracy. Condition one shows that, except for *Info-de-time*, all agents updating from direct experiences provide more accurate appraisals than *Cheat* and *Naive*, which do not update from past experiences. The third condition of the experiment is even more convincing regarding the usefulness of information from experiences. Two information-based agents, one with and one without updating from direct experiences, were tested in the same condition. The agent that updated from direct experiences had a significantly larger final client share and therefore must have produced more accurate appraisals. Thus, the expectation that updating from direct experiences improves the appraisal accuracy is supported by the experimental results.

For evaluating updating from forgetting, the first two test conditions can be examined. Here two information-based agents updating from direct experiences, one of them also updating from forgetting, were tested in the same condition. In the condition with the agents *Cheat* and *Naive*, the agent *Info-de* scored better than *Info-de-time*, but the difference is not significant. In the condition with the agent *Changing*, the agent *Info-de-time* updating from forgetting, has a significant larger client share than *Info-de*. This supports the expectation that updating from forgetting would contribute to more accurate appraisals.

The last type of information, updating from reputation information, has been examined in the third condition. The participating agents are the information-based agent to be evaluated, combined with the three test-agents *Cheat*, *Naive*, and *Providing* which

	Cheat		Naive		Providing		Agent	
	Bank	*Client*	*Bank*	*Client*	*Bank*	*Client*	*Bank*	*Client*
info-de-time	43252	23.1	12986	10.6	34889	23.3	34245	22.7
info-rep-time	45337	22.3	15363	12.7	35337	23.5	28713	21.1
info-de-rep-time	41076	21.3	14089	10.8	34988	23.4	35099	24.5

Table 3. Averages for three information-based agents in the third set of conditions.

provides reputation information. The agent *Providing* performs very well, so the reputation information it provides is supposed to be useful. Agent *Info-rep-time* does not update from any of its own experiences, so its performance only depends on updating from reputation information. *Info-rep-time* ended with much larger client shares than *Naive*, so it seems to use *Providing*'s reputation information profitably. This observation supports the expectation that the use of reputation information would increase the average appraisal accuracy of an information-based test-bed agent. Of course this conclusion only holds when there is at least one agent in the game that is able and willing to provide useful reputation information.

The results show that all three types of updating contribute to appraisal accuracy, but do they also work well in combination? Updating from forgetting can be used in combination with the other two types of updating without hindering them. However, updating from information from direct experiences and from reputation information cannot be added to each other. When more reputation information is used, less information from direct experiences can be used and vice versa. The results show that in both condition one and three, the use of all available types of information yields the most accurate appraisals.

However, in the first condition *Naive* is the only agent providing reputation information and it assumes that each agent is trustworthy, so it always provides reputations with the value 1. So the good performance of the agent using reputation information in this condition cannot be due to its updating from reputation information. In the third condition however, useful reputation information is provided and the agent *Info-de-rep-time* seems to make good use of it. So the results support the expectation that all three types of updating contribute to providing more accurate appraisals, and the information-based agent using all three types of updating provides the most accurate appraisals.

The experiments performed are not exhaustive and when interpreting the results, some remarks should be kept in mind. First, an agent's performance depends a lot on the other participants in a test-bed game. For example, an agent with a very sophisticated model for dealing with reputation information only profits when other agents are prepared to provide reputation information. A cooperative agent functions very well with other cooperative participants, but it might perform very badly with non-cooperative participants. In the experiments, four test-agents were used, *Naive*, *Cheat*, *Changing* and *Providing*, which show quite simple and obvious behavior. The use of more complex test-agents would provide more information. Moreover, conditions with larger numbers of participants would create new situations and might yield extra information.

A second consideration is the choice of the ART test-bed. A general problem of all test-beds is *validity*: does the system test what it is supposed to test? Especially when complicated concepts are involved, it is difficult to prove that a test-bed just examines the performance of a model on that particular concept. The aim of the ART test-bed

is to compare and evaluate trust- and reputation-modeling algorithms [4]. But what do the developers exactly understand by trust and reputation? The ART test-bed is quite complicated and allows so many variables that it is sometimes difficult to explain why something happened.

A final remark about the experiments is that in the translation of the trust model to a test-bed agent some adjustments and adaptations had to be made. Not every part of the model can be used in the ART test-bed. Sierra and Debenham's model [8] allows updating from preferences and different power relations between agents; these facets cannot be tested by the ART test-bed. On the other hand, the trust model lacks theory for some topics needed in the ART test-bed. The updating from reputation was not very elaborated in the model [8] and had to be extended. Besides, the information-based trust model does not provide a negotiation strategy: it is a system to maintain values of trust. The strategy used might have influenced the test results.

8 Conclusion and further research

The goal of this article is to examine Sierra and Debenham's information-based model for trust [8]. Therefore, an agent based on the model has been implemented and several experiments in the ART test-bed have been performed. The experiments showed that the information-based agent learned about its opponents during a game session and could distinguish between cooperating and non-cooperating agents. They also demonstrated that the three examined types of updating (from direct experiences, from reputation information and from the evaporation of beliefs as time goes by), all improved the agent. So in general expectations have been met: the results are promising and the information-based approach seems to be appropriate for the modeling of trust.

The diversity and the amount of the experiments could be extended. The information-based agent could be tested in more conditions with different test agents and with larger amounts of participating agents. It would also be interesting to pay more attention to the agent's strategy. Besides, the implementation of the agent could be improved. Some aspects of the trust model could be translated more literally to the implementation of the information-based agent. Even another test-bed could be used, as the ART test-bed is not able to evaluate all aspects of the theory. All these suggestions would deliver new information about the model and would justify making stronger statements about it.

As to Sierra and Debenham's trust model itself [8, 9], its core seems to be robust and clear: they use a clear definition of trust and probability distributions are updated from a set of beliefs with the principle of minimum relative entropy. The experiments support the model. To further improve it, more work could be done on other concepts related to trust. For example, now it provides some initial ideas about how to deal with reputation and other types of social information. But social aspects are becoming more and more central in the field of multi-agent systems lately, so a contemporary model of trust should give a complete account of it. So, it can be said conclusively that the core of the model seems to be a good approach, but for a fully developed approach to trust and reputation more work should be done. This should not be a problem, because the model is flexible and provides ample space for extensions.

Acknowledgements. Carles Sierra's research is partially supported by the OpenKnowledge STREP project, sponsored by the European Commission under contract number

FP6-027253, and partially by the Spanish project "Agreement Technologies" (CONSOLIDER CSD2007-0022, INGENIO 2010).

References

1. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review **24** (2005) 33–60
2. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision Support Systems **43** (2007) 618–644
3. Mui, L., Mohtashemi, M., Halberstadt, A.: Notions of reputation in multi-agents systems: a review. In: AAMAS '02: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems, New York, NY, USA, ACM Press (2002) 280–287
4. Fullam, K., Klos, T., Muller, G., Sabater, J., Topol, Z., Barber, K.S., Rosenschein, J.: A specification of the agent reputation and trust (ART) testbed: experimentation and competition for trust in agent societies. In et al., F.D., ed.: Fifth International Conference on Autonomous Agents and Multiagent systems (AAMAS-05), Utrecht, The Netherlands (2005) 512–518
5. Ramchurn, S.D., Huynh, D., Jennings, N.R.: Trust in multiagent systems. Knowledge Engineering Review **19** (2004) 1–25
6. Castelfranchi, C., Falcone, R.: Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In Demazeau, Y., ed.: Proceedings of the Third International Conference of Multi-agent Systems (ICMAS98). (1998) 72–79
7. Sabater, J., Sierra, C.: REGRET: reputation in gregarious societies. In: AGENTS'01: Proceedings of the Fifth International Conference on Autonomous Agents, New York, NY, USA, ACM Press (2001) 194–195
8. Sierra, C., Debenham, J.: An information-based model for trust. In et al., F.D., ed.: Fifth International Conference on Autonomous Agents and Multiagent systems (AAMAS-05), Utrecht, The Netherlands (2005) 497–504
9. Sierra, C., Debenham, J.: Trust and honour in information-based agency. In Stone, P., Weiss, G., eds.: Proceedings Fifth International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2006, Hakodate, Japan, ACM Press, New York (2006) 1225 – 1232
10. Sierra, C., Debenham, J.: Information-based agency. In: Proceedings of Twentieth International Joint Conference on Artificial Intelligence IJCAI-07, Hyderabad, India (2007)
11. MacKay, D.: Information Theory, Inference and Learning Algorithms. Cambridge University Press (2003)