Improving the ART-Testbed, thoughts and reflections

Mario Gomez[1], Jordi Sabater-Mir[2], Javier Carbo[3], and Guillaume Muller[4]

[1] Computing Science, University of Aberdeen, AB24 3UE Aberdeen, UK
m.gomez@abdn.ac.uk
[2] IIIA, CSIC, Campus UAB, 08193 Bellaterra, Catalonia, Spain
jsabater@iiia.csic.es
[3] GIAA, Carlos III University of Madrid, Av. Universidad 30, Leganes 28911, Madrid, Spain
jcarbo@inf.uc3m.es
[4] LORIA/INRIA Grand-Est, Campus Scientifique - BP 239 - 54506 Vandoeuvre-les-Nancy CEDEX, France

Abstract. The Agent Reputation and Trust (ART) Testbed initiative was launched with the goal of establishing a testbed for agent reputation- and trust-related technologies. This initiative has lead to the delivery of a flexible/modular prototype platform to the community. After two years and four successful competitions, several issues have arisen that require some attention to maintain the focus of the testbed as a research oriented tool as well as the interest of the participants. This paper presents these issues and suggests some possible modifications that we think can improve the testbed.

1 Introduction

The accomplishment of complex tasks in MultiAgent Systems often requires agents to collaborate. However, it is impossible for agents to assume that the others will cooperate sincerely. Establishing collaborations therefore puts the agents at risk. To reduce the risks, an agent can rely on the trust/reputation it assigns to others. Thus, by selecting its partners, it can avoid unreliable agents and select partners that have the best potential to fulfill their duties.

In recent years, many trust/reputation models have been proposed. It is very difficult to compare their respective performances as many application domains and metrics have been utilized. Researchers [**?**,**?**,**?**] have recognized that objective standards are necessary, for the public to be provided with transparent evaluations.

Based on enthusiastic response from the agent trust community, the Agent Reputation and Trust (ART) Testbed initiative was launched in 2004, and has delivered a prototype testbed platform for agent trust/reputation models [**?**,**?**] in 2005. With this prototype, two competitions were organized in 2006, one in Madrid, Spain, and the other associated to the AAMAS international conference in Hakodate, Japan. Based on the success of these competitions, both

competitions have been repeated in 2007, in Valencia, Spain and Honolulu, USA respectively.

During these two years when the prototype has been utilized by several researchers, the ART-testbed members have gathered feedback from the agent trust community. This paper presents a synthetic analysis of this feedback and proposes new directions of work.

In Section 2, a brief description of the scenario and functioning of the testbed is given. Section 3 discusses the most prominent subjects of feedback and draws up the plans for both the future of the platform and the competitions. Finally, Section 4 concludes on the experience acquired during these two years of development and use of the prototype.

2 The Art-Testbed

The ART-testbed platform can be used in two modes: experimentation or competition. In this paper, we focus on the competition mode. In this mode, the parameters are decided by the competition organizers and each team of researchers provides a unique class of agent. According to the rules fixed for a competition, one or several instances of this agent classes are put into the competition with those provided by the other participant teams and a certain number of dummy agents provided by the organizers.

The scenario defined for the ART-testbed platform takes place in the art appraisal domain. In this section, we give a quick overview of this scenario. For reasons about the choice of the scenario and for a more detailed presentation of the scenario and the platform, please refer to [**?**].

Agents, which are implemented by researchers based on a provided schema, function as art appraisers. The scenario is so designed that agents are globally in competition for client share, but able to cooperate on single appraisals.

At the beginning of a game, the platform assigns to the agents varying expertise degrees in different artistic eras. The expertise is modeled as the level of accuracy generating appraisals. The expertises are distributed fairly among the participant agents. The agents' expertise can be fixed for an overall game or change several times during a game.

A game is organized in turns, that are iterated a customisable number of times. A typical game turn spans as follows:

- For a fixed price, clients (which are managed by the platform) ask appraisers to provide appraisals of paintings from various eras. The paintings are considered to have a unique market value, that is only known by the platform.
- The appraiser agents pay the platform to compute their appraisal of the paintings. This step is used (1) to prevent agents from using diverging strategies of appraisal (that would interfere with our aim to evaluate only the agents' ability to model trust/reputation) and (2) in order to enable agents to pay more to get a more accurate appraisal (in the limits of the expertise of the agent). The latter being associated with the fact that an appraiser

may wish to spend less resources appraising a painting, at the expense of getting a less accurate appraisal.
- If an appraiser is not very knowledgeable about a painting that has been submitted by one of its clients, it can purchase "opinions" from other appraisers. An opinion is the estimated value of a specific painting by one appraiser agent. Since an appraiser to which the opinion is requested is not necessarily fully competent, nor sincere for the given painting, its appraisal can be inaccurate. There is therefore a specific protocol that an appraiser agent uses to request an opinion. During this protocol, the requester gets an estimation of the expertise of the potential opinion provider (provided by the potential opinion provider itself) and decides whether to continue its purchase of an opinion or abandon it with no fee.
- Appraisers can also buy and sell "reputation" information about other appraisers. This is done by exchanging reputation weights. Reputation weights are values in [0,1], 1 being the highest reputation. Providers of reputations have to convert values from their own internal model representation into the standard interval used for reputation weights before sending them. Requester agents have to convert them back from this standard interval into they own internal representation.
- Appraisers ask the platform to compute their final estimations for the paintings. This is also done by the platform to guarantee that all agents use the same method.
- At the end of the turn, agents' appraisals are compared with the true values of the paintings stored by the platform. Error in the estimations can be computed and are used to update the client share repartition: appraisers whose appraisals are more accurate receive larger shares of the client base for the next turn. Then, agents are informed of the true values of the paintings and they can update their trust/reputation models.

During a competition, appraisers compete in several games. The winner of a competition is determined by a computation based on the agents' bank balances at the end of each game. The agent that sums up to the highest bank balance wins.

According to the description given above, an agent gets earnings in three situations: when clients submit paintings to appraise; when it sells opinions; when it sells reputations. Agents spend money when they buy opinions or reputations. As the bank balances of the agents are not limited in the negative values, an agent can always buy opinions or reputations. This definition of the scenario guarantees that agents earn more if (1) they manage to become trusted providers of opinions and reputations, (2) they manage to become good appraisers: they are able to learn as quick as possible to identify agents that can provide them with good opinions/reputations and (3) they adapt quickly their models in cases where the other agents change their behaviors.

The main metric of the testbed is the bank balance, since it is used to define who is the winner. However, there are other parameters in the games that can be observed and analyzed by the researchers to compare the performance of the

agents: the transaction fees (money earned from clients or earned/payed from/to opinion or reputation providers); the reputation weights (for each other agent in the system); the average appraisal error (defined by the standard deviation of the final opinion of an appraiser to the true values of the paintings); the transactions counts (number of opinions/reputations sold to or purchased from another agent); the complete lists of all the messages exchanged between agents.

3 Issues and suggested improvements

3.1 Allocation of expertise values

A crucial element of the ART testbed is the notion of expertise. Participant agents (appraisers) are assigned different expertise values for a number of artistic eras. To ensure that agents are different, these values are obtained randomly; each expertise value s^* is calculated using an uniform distribution in {0.1, 0.2, ...,1}. The opinions of an agent on the true market value of a painting are generated by the simulator in such a way that opinion errors adhere to a normal distribution with mean equals to 0 and a standard deviation s defined as:

$$s = (s^* + \frac{\alpha}{C_g})t_k \tag{1}$$

where s^* is the *expertise level* of an agent for a particular era, C_g is the money invested by an agent to generate an opinion, α is a parameter that defines the influence of C_g on the error, and t_k is the true value of a painting, its market price.

From the definition of s follows that the so called expertise level s^* is actually the minimum standard deviation of opinions' error, and thus, $s^* = 0.1$ constitutes the best case (the most expert an agent can be), since it implies a narrow distribution of errors, i. e., more accurate opinions.

Agents are motivated to obtain accurate appraisals to increase their client share, thus if an agent is not expert in an era, it has good reasons to buy opinions to other agents that are more expert in that era. Complementarily, if an agent is expert in an era, it is not worth to buy opinions to other agents, because it has both a cost and a risk (an agent can provide a bad opinion deliberately). That means that agents that are receiving good expertise values (low $s*$) from the sim have a clear advantage over those agents that get bad expertise values (high $s*$). As a consequence, the assignment of expertise values at the beginning of a game can have strong impact on the relative performance of the participants.

All in all, since expertise values are given randomly, and the number of eras is low, the relative performance of a group of agents over a number of games tends to be quite unsystematic, showing a high variance.

Figure 1 illustrates this problem. This figure shows the final earnings of the five finalists in the 2006 ART International Competition, along 10 games. We can see that the results of each game are quite different. Even the average results are so much adjusted that in many cases we cannot establish whether an agent

performs better than another (there are no statistically significant differences); in particular, the second, third and fourth positions (Neil, Frost and sabatini) are too close as to generalize the results and state than one agent is better than another.

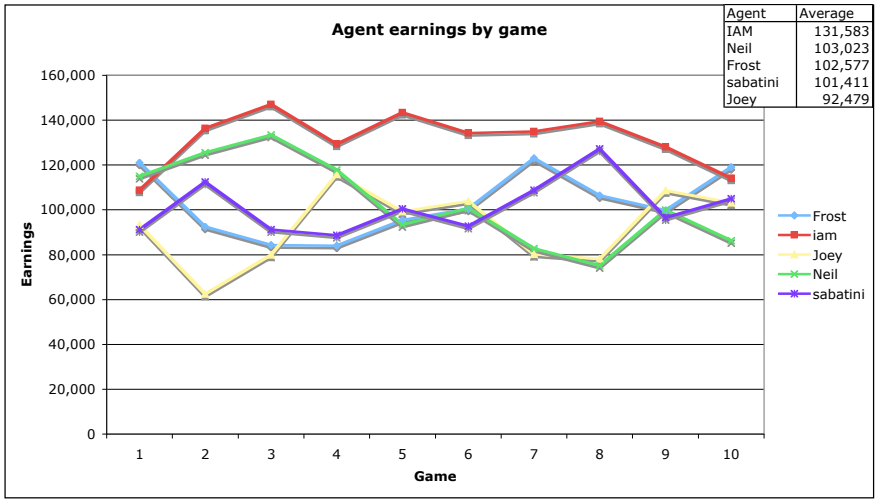

Fig. 1. Final games in the 2006 ART International Competition

As a consequence of the random allocation of expertise values, many experiments are required to draw reliable conclusions. We have verified this through extensive experiments with the first version of the ART. To palliate this problem, we have introduced two new strategies to assign expertise values in a fairer way, ensuring that all the agents have the same average expertise.

In the first strategy (R1), for each agent, the testbed assigns the expertise values s^* for half of the eras randomly, and the other half is assigned using complementary values $1 - s^*$. As a result, agents may have completely different expertise values, but the average expertise is guaranteed to be equal to the theoretical average of the uniform distribution, which is half the range of the variable, that is 0.45 (1-0.1/2). Here, we are controlling only the average expertise.

In the second strategy (R2), we start with a single list of expertise values, and then assign the same list of values to every agent, but shuffling the elements of the list (changing the order of the elements randomly). Therefore, all the agents have the same average expertise and the same precise values, but assigned to different eras. In this case, we are controlling not only the average expertise, but also the variance.

Figure 2 shows the earnings of the finalists in the 2007 ART International Competition, this time using the second of the new expertise allocation strate-

gies (R2). In this case, participant agents were competing also against a mix of dummy agents included by the organizers of the event: 5 "nice" dummies, 5 "neutral" dummies, and 5 "bad" dummies). Furthermore, three experimental situations were considered: A) No expertise changes per game; B) One expertise change per game; C) Two expertise changes per game. We can see in the figure that now the relative positions of the participant agents are much more stable, specially if we look separately at the three experimental situations: In situation A, all the agents have finished in the same position: IAM, Jam, Blizzard, Spartan and ZeCariocaLes; and in situation C there is only one variation, in game 4.

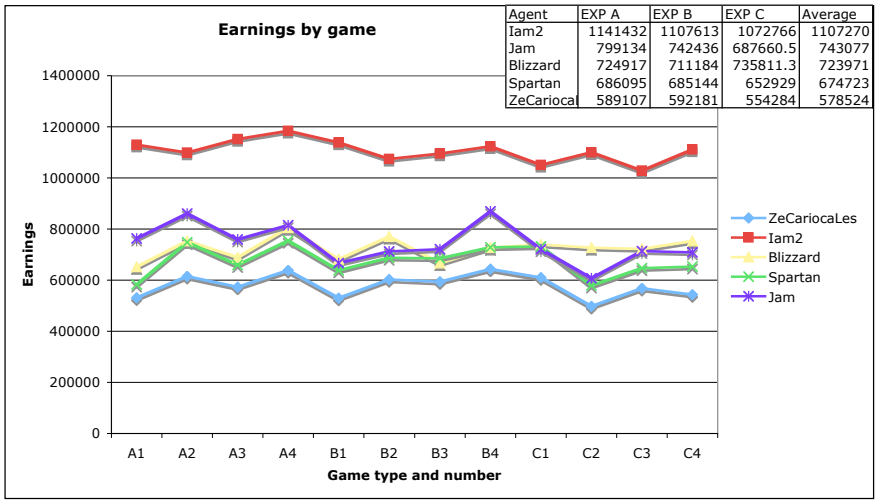

Fig. 2. Final games in the 2007 ART International Competition

We conclude this section with some notes about potential pitfalls derived from the fairer expertise allocation strategies. First of all, by controlling the allocation of expertise values, we are reducing the number of situations an agent may face, so we are losing the opportunity of testing the behavior of an agent under extreme situations, like those involving an agent with a very low average expertise. Although this seems a bad situation to compare agents in a competition, it may still have interest from an experimental point of view. The ideal solution would probably involve a number of games in which a given agent is selectively put in such an extreme position, ensuring that in the average all the agents are exposed to the same situations.

A second pitfall is associated with the allocation strategy R2: by making all the agents having the same distribution of expertise values, we reduce the variability, and thus the accurate assessment of the expertise of other agents is not so important (Section 3.2 discusses this issue more deeply). Furthermore, if

an agent knows that this strategy is being used, it could exploit the knowledge of its own distribution of expertise values to estimate the expertise values of other agent without requiring additional information, just by probability[5].

3.2 Low discrimination between similar models

The main factor in determining which agent wins is the client share, that depends on the accuracy of one agent's final appraisals.

$$p^* = \frac{\sum_i (w_i . p_i)}{\sum_i (w_i)} \tag{2}$$

where p^* is the final appraisal for a painting, p_i are the opinions of different agents, and w_i is the weight assigned by an agent to each of the opinions.

An agent can use its own opinions to compute the final appraisal. Due to the way opinions and appraisals are computed, the error tend to decrease when adding more opinions, because positive errors compensate for negative errors. Since other agents may cheat about their real expertise values, and an agent has perfect knowledge of its expertise values, then an agent can reduce the risks of producing very inaccurate appraisals by including their own opinions in every appraisal, even for those eras in which it has low expertise. The best example of the successful use of this strategy is provided by the winner agent of the 2006 International Competition [**?**].

Although the use of self opinions is meaningful, and seems appropriate at first glance, it introduces some problems when trying to compare and assess a trust model. Due to the high impact of self opinions over final appraisals, it is very difficult to discriminate between different trust models, specially if these models are only slightly different, as use to be the case when doing experiments. The difficulty to discriminate between similar models has been reported in [**?**]: the authors compared three approaches to handle the discrepancy between the quality of opinions promised by an agent, and the quality actually found. Extensive experimentation was unable to find significant differences between the three models. However, the authors were finally able to discriminate among the models by denying agents the use of their own opinions. This issue could also be one of the reasons explaining the small differences in average performance among the agents that participated in the 2006 and 2007 competitions (see figures 1 and 2).

We are currently considering several options to reduce the impact of self opinions to increase the ability of the testbed for discriminating among different trust models. Two of these options are introduced below.

- A first solution is suggested in [**?**]: to deny the use of self-opinions. By doing that the computation of trust, which is used to calculate the weights in

[5] To understand this, consider the extreme situation in which an agent knows another agent in all eras but one, then it could infer the expertise in that era simply by comparing both distributions of expertise values (the other's and itself's)

equation 2, would have a stronger impact on the final appraisals, thus helping to reveal the differences among different trust models.
- A second solution is to deny agents the knowledge about their expertise values. In such case, an agent would have to assess its own expertise by requesting opinions and learning from them, as it does with other agent's opinions. However, once the expertise values are estimated, an agent can have total confidence on its intentions, so this solution will probably not work after a number of rounds.

Another element that currently could be limiting the discrimination power of the testbed is the narrow and limited range of expertise values considered; currently, there are only 9 possible values: {0.1,0.2, ..., 1 }. This has two undesirable consequences: first, the quality of the opinions provided by different agents are moderately different, so this is not so important to detect small differences; and second, it is easy to find agents with the best expertise value, thus agents have low motivation to know other agents, which is an additional reason for not favoring the use of reputation information (see §3.5)

We are considering two measures to palliate the negative consequences of the narrow and limited range of expertise values:

- On the one hand, if we allow a wider range of values, then the decisions about which agents to choose as opinion providers will have a higher impact on the accuracy of final appraisals.
- On the other hand, if we allow more expertise values, then it will be harder for an agent to assess trust (which implies that the trust model should be more accurate). In addition, it will be harder to find agents with the best expertise, which in turn will motivate agents to gather information about many more agents (e.g. using reputation information).

We still have to make experiments to test whether these measures are really effective. Another thing to take into account is that some of these measures may collide with the strategies discussed in Section 3.1. Specifically, the second of the new strategies for allocating expertise values (R2) would probably reduce the positive impact of enabling more expertise values, because it ensures that all the agents have the same distribution, which clearly goes against the idea of increasing variability among agents.

3.3 Measures of trust

One of the potential confusions in the ART testbed is that the focus may be on how money is invested much more than on how knowledge is used. For instance we can observe how one of the points of the success of agent IAM [**?**] in the competitions was the search of the most profitable quantity to invest in opinions according to the corresponding equation.

Furthermore, once some agent has earned some more money than others, (even if it is acquired by a better use of trust knowledge), this agent may obtain better results exploiting the money rather than trust knowledge. So the

results of the next iterations would be then non-conclusive related to the goal of comparing the use of trust knowledge. That economic bias is also supported by the fact that agents are simultaneously clients and providers of opinions. Since the impact of an agent performance as a provider influences the global perfomance of an agent, the results are not measuring just the ability to assess other agents. This feature could be desirable for some experiments, like those concerning cooperation and coalition formation or social networks, but this is not so good for those experiments willing to compare trust models (such as the ART Competition). Even more, we could state a difference between trust models and trust strategies. Trust models would involve just how trust is computed, in other words how it is updated from direct experiences and indirect opinions, and nothing else. On the other hand, trust strategy involves how trust is applied in the communication decisions. Obviously they are strongly related but still different issues. Since one strategy from one source/author and one model from another source/author may be easily combined, it would be interesting that they can be separatedly evaluated.

So we can outline three categories of challenges to compare trust models:

1. Independency from economic/mathematical formulations
2. Single-roled agents: either client or provider.
3. Evaluating models and strategies separatedly.

Although many measures may be compared in ART testbed (such as average appraisal error, efficiency, performance, stability, ...), just points (representing money) were the chosen measure in the ART competitions since as ART testbed is defined, points represent the global performance of an agent in the game better than the other measures. But this is one of the points of controversy of current ART testbed: is a money-based comparison of trust models fair?

Among other consequences, the central role of economic affairs in the current testbed remarks the importance of the cost/benefit analysis on mathematical formulations. Achieving a desirable complete absence of bias due to equations is imposible but, for instance, a more direct and proportional outcome from money invested could avoid a misplaced focus of research pointed out by IAM authors [**?**]. In other words, changing an exponential curve by a linear function would make at least more difficult to exploit any the search of an optimum investment amount in cost/benefit analysis. This study of the better amounts to be chosen was carried out by all the competitors in one way or another. This is a waste of research time, since this curve was not supposed to be the key of any trust model.

Additionally to this linear function, a more restricted possibility of investment would be helpful to approach to this goal, for instance limiting the minimum and maximum amount that any agent may invest in an appraisal. This allowed range would make more difficult to exploit the different early earnings into very different knowledge acquisition. On one hand, a perfect valuation has to be imposible to achieve (this is rightly modelled currently with ART testbed) but on the other hand, a null investment (as current ART testbed allows) makes very little sense in this research context. If no cooperation at all was the chosen

strategy, giving nothing for the money received is not a deceiptful behaviour, it should be then considered as an objective failure of a trade agreement that should be punished by norms and legal autorities, not by a loss of reputation. Personal trust models makes sense when we face subjective evaluations of behaviour, this all-or-nothing approach does not belong to agent-based reputation models, it belongs to law and norm authorities.

These two suggestions do not imply more than slight changes to the testbed definition, in fact current competitor agents could run, just as they are, with a different curve, and a transformation function applied to the invested amounts.

A more radical view that could be also interesting consists of avoiding completely the use of money itself. It would imply then to fix in advance the number of agents that could be asked for help both in appraisals and reputation queries, and to fix in advance the total amount of resources to invest in appraisals (now the points represent 'time' rather than 'money'. Both numbers (appraisal and reputation queries and time to invest in appraisals) would be the same for all the agents in every iteration. Then, this new testbed definition would lead to agents competing on equal terms each and every iteration. The only difference would be then the knowledge about others that each agent would have, and how they apply it in the final weights to compute their own appraisal. The consideration of an alternative testbed as described, could be a more 'pure AI' approach than the current one.

This non-economic view of trust models comparisons, leads easily to the next challenge mentioned above: Should agents be single-roled or they should act jointly as clients and providers. This point is even less clear than the previous one. On the one hand, agents in ART testbed should act as clients and as providers at a time, and this be right in several real-world reputation scenarios. But on the other hand, are providing behaviours relevant to compare trust models? This question drives us to another one: what should we consider precisely as a trust model? If a trust model implies how much trust we assign to the other agents according to their behaviour, then there will be no provider role at all and evenmore, since trust model fits better with a role of evaluators rather than a role of client. This may be a basic definition of trust model, so we can open the definition of trust model considering also how this assigned trust is applied. Then this extended definition of trust model should include decisions such as who to ask for opinions, and not just how much these opinions are trusted. Now trust model of agents involve acting as clients and not only as evaluators, but again there is no provider role in this approach of what a trust model should be. Finally there is another level of trust decisions that agents may take: how trust is applied when we have to answer, when are asked for providing opinions. That is what could be called 'trust strategy'. The point to remark here is the fact that the trust decisions of agents involve not only how good an agent is assessing trust, but also his policies/strategies as a provider to other agents, which has nothing to be at all with the (basic or extended) definition of trust model These three different approaches to trust model definition (basic model: just evaluating, extended model: evaluating and asking, strategy: evaluating,

asking and answering) should be considered in an isolated way since maybe the way trust is acquired was better in one model, while trust-applied-to-questions is better in another one, and so on. In fact, a very good model fooling other agents could hide a very bad model acquiring knowledge about others. Therefore, mixing the three of them could not be the better way to discriminate the best model of each type of trust decision.

What is then the magic solution? Rather than playing three different games, the real goal should be how to discriminate (and compare) the three of them (one by one) in the same game. This should be the point. Since this is not easy, perhaps a variant of this goal may be a right approach. Same agents, same testbed, but different scenarios. In one of them all act as clients and providers applying a complete trust strategy (with asking, answering and acquiring knowledge issues), this is the current ART testbed approach. Another one, were competitors are never asked for help (neither appraisals and reputations), other predefined agents (from the testbed) act as the only providers in the game, and finally a third scenario were competitors just provide the weights of each provider to the testbed, and then it will be the testbed itself who asks other agents in behalf of the competitor agent, according to the trust model (in other words, the weights). This last scenario implies that no asking neither answering decision relies upon competitors.

3.4 Changes in the environment

Since dynamism is quite important in real environments (e.g. markets), the current approach of ART Testbed is not appropriate, since it doesn't resemble the dynamics of real systems, such as evolving prices in a market, evolving expertise with continuous practise and the appearance of new type of products.

The evolution of prices in the market is a very well-known issue that leads to inflation and deflation trends that may increase the interest of economics research community in ART Testbed. Perhaps some economist research should join the ART developers group to include these economic issues in the environment. This way of opening testbed to other research communities, is a path that may lead to a major success of the ART testbed. For instance if we could define the behaviour of several agents in terms of action operators over prefixed scenarios to be aplicable on ART testbed, then Planning community may be interested. Additionally, machine learning community could be interested in ART testbed if data of previous games could be used in advance to apply data minig to them and see how this AI technique may produce a competetive agent in this testbed.

The second issue was how should expertise evolve along time. Of course, it should be a progressive modification rather than a sudden one (to be realistic). These progressive changes, adding/substracting very small quantities should make some sense (not to be random). The best way to justify these changes is imitating real world. The one who applies expertise continously, improves expertise, the one who does not, loses expertise. These evolution of expertise would lead to interesting new trust strategies in ART testbed, since agents may have new goals such as becoming overspecialized in some eras, or trying to be an

all-era agent. This way of changing expertise is more realistic, and implies new challeges to competitors rather than the current ART testbed expertise changes (randomized but keeping the average expertise, or applying the same overall modification over all the agents).

Finally an important issue is the appearance of new type of products (in the ART Testbed world, they would be new eras). If the approach taken when a new era arises is giving minimum expertise for this new-born era to all agents, then acquiring expertise through practice becomes even more important. Therefore agents will start a race to become the first to know about this new era. An alternative approach would be to apply new expertise according to similarities between old and new eras. This would increase the complexity of maintaining a balance among the given expertise of all competitors, and it will not imply a race to enter into a new market on equal terms.

3.5 Motivating the use of reputation

One thing that is obvious after analysing the results in both the first and the second competition is that participants are barely using information based on reputation. The group of the University of Southampton, winners of the two competitions, leave it clear in the article [**?**] that describes its IAM agent: "the IAM agent does not request reputation values from other participants ... it relies solely on the variance estimator to estimate the performance of an opinion provider". The reasons argued for that decision were the few number of agents, that makes the use of direct experience enough to know the expertise of the other partners, and the difficulty to interpret the meaning of the reputation exchanged value given there is not a common ontology. After detecting this problem in the first competition, some changes were proposed for the second one, the most important the increase in the number of players by adding dummy agents and the change of players expertise during the game. The idea behind these changes is to make more difficult for a participant to know what is the expertise and behabiour of a possible partner. In other words, to force participants to use reputation because the search space is too big and volatile to rely only on direct experiences. However these changes were not enough to modify the behaviour of the participants in the second competition, being again the use of reputation something marginal. In fact, the five agents that arrived to the final during the second competition were not using reputation exchange at all. A further analysis arised another problem. As the game is designed now, agents can ask for an appraisal from another agent that will send back a level of certainty based on its expertise in the specific era of the painting to be appraised. This information is intended to be used by the agent to decide if it is worth it or not to go ahead with the appraisal request. The information about certainty is free, and given that asking for an appraisal can be aborted with no cost after receiving the certainty value, asking for appraisals just to get the certainty is a better (and cheaper) method to know the expertise of another agent than using reputation information. The game do not impose any restriction on asking for appraisals to all the other players just to know their expertise in a specific era. Of course,

they can lie when giving the certainty values but this problem is also present in the exchange of reputation information.

Here there are some possible changes that we think could help to motivate the use of reputation. As demonstrated during the second competition, it is necessary to use a combination of several of them to be really effective.

- *Increase the number of participating agents.* In the second competition the total number of appraisers (participants plus dummy agents) was 20. This number is still too low to justify the use of reputation. Actually, reputation is useful in medium-big societies where the use of direct experiences as a single source of information is not worth it (either because is difficult to get this direct experience or because it is too expensive). Of course here there is a problem of performance that has to be solved and that was the main reason for not increasing the number of agents during the competition.
- *Increase the cost of direct experiences.* Currently, the cost of an opinion is already quite high with respect to the cost of reputation information so taking into account the problem about "certainties" should be enough.
- *Use a common ontology that gives meaning to the reputation value range.* The use of a common ontology in the testbed is something necessary to give a semantic to the reputation values. The work presented in the article [**?**] goes in that direction by integrating the ontology of Casare and Sichman [**?**] with the ART testbed.
- *Change the expertise of the agents during a game.* (see also section 3.4) This, together with the increase of the number of participants, is intended to make the environment less predictable. However, the way it was implemented in the second competition where several times during the game all the agents changed their expertise at the same time, does not give the expected results. The expertise change has to be asynchronous among agents to avoid that the fact your expertise is changing be a signal that the expertise of the other agents also have changed.
- *Remove the use of certainties from the opinion request protocol or pay for certainties with a cost similar or higher than reputations.* We have explained before the problem of exchanging certainties as part of the appraisal request procedure. One possibility is just to remove this step in the protocol. Another possibility is to charge the agent a fixed amount of money that should be bigger than the cost of reputation information but less than an opinion.
- *Add certainties about reputations.* As we have seen, in the current version of the ART, the agents that want to buy an opinion first receive a certainty value representing the degree of confidence the possible opinion provider has on the specific era. One possibility would be to add a similar step in the reputations transaction protocol so as to allow an agent to say whether it knows or not another agent. If agent A is requested by B about C, then A could say B that he does not know C (by providing a low reputation certainty), so B shouldn't expect a good reputation from A. As it is now, an agent cannot tell another agent that he doesn't know a third agent and has to provide always a value. Adding this into the protocol would increase the reliability of reputation values and make them more valuable.

4 Conclusions

In this paper, we have discussed some aspects of the ART testbed that pose a challenge to utterly accomplish its goals. Furthermore, we analyze some issues that currently limit the applicability of testbed as both an experimental and a competition platform.

First, we have discussed the problems that arose from the randomized allocation of expertise values, which often resulted in unfair situations involving some agents with better expertise values than others. This makes necessary to repeat experiments many times in order to generalize the results, and was a handicap to obtain conclusions from the few games that are performed during a competition. In Section 3.1 we have described new expertise allocation strategies that are intended to deal with the former problems. However, these strategies introduce some side effects that may lead to game strategies in competitions.

In Section 3.2 we have introduced additional issues that make it difficult to compare trust models. Specifically, we have pointed up the use of self opinions and the limited range of expertise values as two of the reasons underlying this problem. We have briefly commented some possible solutions to deal with these problems, based on a wider range of expertise values and denying the use or reducing the impact of self-opinions.

Next, we have discussed in section 3.3 which should be the right way to measure the performance of agents in the ART testbed. Specifically we pointed up the excessive importance of economic/mathematical formulations, and we suggest the possibility of evaluating trust models and trust strategies separatedly using restrictive scenarios where agents could act either as clients or providers.

Afterwards, Section 3.4 have proposed the use of ART environment as a possible application domain to other research communities different from agent communiy such as economics, planning, machine learning as a way to increment the visibility of trust and reputation issues, and ART testbed itself.

The use of reputation information by the participants during the two competitions was far less than expected. Even the changes proposed after the first competition to solve that (increasing the number of agents and global changes in the expertise of the participants during the game) were not enough to promote the use of reputation information during the second competition. It is clear then, that more radical changes complementing those adopted after the first competition are necessary if the testbed has to be a tool for testing reputation models and not simply models that use only direct experiences. In Section 3.5 we have presented a list of these changes that include things like the use a common ontology that gives meaning to the reputation value range or removing the use of certainties from the opinion protocol.

All in all, the high quantity and variety of issues and solutions that we have gathered, make it particularly difficult to address all of them together. On the one hand, some issues have resulted in a number of alternative solutions that shall be implemented and thoroughly evaluated. Rather than finding the best single solution for every problem, it is rather more likely that different solutions would fit better into different scenarios and situations. On the other hand, some

issues considered here might be interrelated in different ways that are hard to fully realize in advance. We should carefully evaluate the relationships between these issues, and analyze how the proposed changes may result in undesired side effects. To conclude, as a result of all the problems and limitations identified, and the many improvements considered, we expect to further develop the ART testbed along multiple lines, and make it more customizable to suit the needs and preferences of different researchers, as well as providing a richer and more flexible environment for the organization of challenging and interesting competitions.

5 Acknowledgments

This work was supported by the European Community under the FP6 programme (eRep project CIT5-028575 and OpenKnowledge project FP6-027253), the project Autonomic electronic Institutions (TIN-2006-15662-C02-01), CICYT TSI2005-07344, MADRINET S-0505/TIC/0255 and CAM CCG06-UC3M/TIC-0781. Jordi Sabater-Mir enjoys a Ramon y Cajal contract from the Spanish Government.