Trust Decision-Making in Multi-Agent Systems

Chris Burnett **Timothy J. Norman**
Department of Computing Science
University of Aberdeen
Scotland, UK
cburnett, t.j.norman@abdn.ac.uk

Katia Sycara
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA USA
katia@cs.cmu.edu

Abstract

Trust is crucial in dynamic multi-agent systems, where agents may frequently join and leave, and the structure of the society may often change. In these environments, it may be difficult for agents to form stable trust relationships necessary for confident interactions. Societies may break down when trust between agents is too low to motivate interactions. In such settings, agents should make decisions about who to interact with, given their degree of trust in the available partners. We propose a decision-theoretic model of trust decision making allows controls to be used, as well as trust, to increase confidence in initial interactions. We consider explicit incentives, monitoring and reputation as examples of such controls. We evaluate our approach within a simulated, highly-dynamic multi-agent environment, and show how this model supports the making of delegation decisions when trust is low.

1 Introduction

Trust has long been recognised as a vital concept in open multi-agent systems (MAS), where agents may be self-interested, diverse and deceptive [Sabater and Sierra, 2005]. Often in MASs, agents must rely on the abilities, competencies or knowledge of others for the fulfilment of their own goals. By relying on others, agents place their own interests at risk, which introduces the need for trust [Castelfranchi *et al.*, 2006]. Consequently, much attention has been given to the problem of allowing agents to make accurate evaluations of the trustworthiness of potential interaction partners. Less attention, however, has been given to the problem of making decisions about who (and whether) to trust, given the risks, rewards and costs associated with trusting.

In this paper, we seek to address aspects of the decision-making problem in multi-agent systems which are highly dynamic. Such systems are characterised by a high rate of population turnover, where the agent population frequently and rapidly changes, and high degrees of social change, where the social structures among agents frequently change. It can be difficult to form stable trust relationships using existing techniques (e.g. [Huynh *et al.*, 2006; Teacy *et al.*, 2006; Jøsang and Ismail, 2002]). Most existing approaches consider trust as a function of direct and reputational (i.e. provided by others) evidence about the behaviours of potential candidates. In dynamic systems, agents may only participate for a short time before leaving, and so it may be impossible to gather sufficient evidence to evaluate the trustworthiness of partners, or opinion providers[1]. Subsequently, agents may be unable to identify trustworthy partners, or make effective decisions about whether to interact. When agents are sensitive to risk, and can choose not to interact, MASs can become 'paralysed'; without interactions, evidence cannot be gathered, trust cannot be formed, and interactions may not take place. This can occur when the risks associated with interactions are high, but trust relationships between agents are weak.

In order to overcome this problem, organisations may make use of controls which permit interaction when trust is low, providing initial evidence from which to bootstrap trust evaluations. In this paper, we will investigate ways in which agents may also use controls to mitigate the perceived risk in initial interactions. We will consider three kinds of controls which can be used in the cases where trust is insufficient:

- **Explicit incentives:** the trustor creates a contract specifying the compensation (in terms of utility) that the trustee will receive, dependant on the outcome.
- **Monitoring:** the trustor expends additional effort/utility in order to observe the behavioural choices of the trustee.
- **Reputational Incentives:** the trustor calculates the reputational gain (or damage) that a trustee will experience as a result of good (or bad) feedback being communicated to the society, and considers this as an additional incentive.

The rest of this paper is organised as follows. In Section 2, we discuss the underlying components of our approach. In Section 3, we discuss the process by which trustors decide which trustee to delegate to. In Section 4, we evaluate a our model within a simulated highly-dynamic multi-agent system, and present our results. We conclude in Section 5.

[1]Throughout this paper, we will refer to agents in these roles as trustees and recommenders respectively. We will refer to agents deciding to trust as trustors.

2 Framework

In this section, we introduce notation for discussing agents, tasks, groups, and outline the assumptions we make about the underlying trust evaluation model

2.1 Agents and Tasks

We assume a society of agents, $A = \{x, y, ...\}$. A trustor $x \in A$ desires to see some task τ accomplished and considers motivating a trustee $y \in A$ to undertake the task on its behalf. Each task τ has two possible outcomes $\mathcal{O}_\tau = \{o^+, o^-\}$, representing satisfactory and dissatisfactory task outcomes (we assume that all trustees share the same task evaluation criteria). Both trustors and trustees possess utility functions which define preferences over these outcomes. We denote the trustors utility function as $U(o)$, and the trustees as $V(o)$. These functions need not be publicly visible. As we assume $\mathcal{O}_\tau = \{o^+, o^-\}$, we refer to $U(o^+)$ as the reward of delegation success, and $U(o^-)$ as one of the risks of failure. We define, for both agents, the minimum expected utility (EU) that the agent requires in order to consider interaction worthwhile. Let $U(abs)$ and $V(abs)$ denote the utility that trustor and trustee (respectively) can obtain by abstaining. These values are visible to both agents. Where multiple trustors and trustees are concerned, we will denote the relevant agent with subscripts, for instance: $U_x(abs)$ and $V_y(abs)$.

Delegation often involves accepting that the delegated agent will have some degree of autonomy in selecting the method by which a task will be carried out. To model this, tasks are associated with a number of effort levels $\mathcal{E}_\tau = \{e_1, \ldots, e_m\}$, each of which incur a different cost $Costy(e_j)$ for y and result in a different probability distribution over O_τ. This formulation can be used to capture many different notions, such as levels of exertion or investment in the task on the part of y. It may be useful to consider effort levels as alternative, specific, methods of carrying out some higher level goal τ. We assume here, without loss of generality, two effort levels, $E_\tau = \{e^+, e^-\}$, representing *high* and *low* effort.

When delegating, the trustor may devise a payment function, or *contract*, which specifies how the trustee will be compensated, depending on some observed outcome of the delegation. In general, this takes the form of a function $C^x_{y:\tau} = \mathcal{O}_\tau \rightarrow \mathbb{R}$, where y is the agent to which the contract applies, which determines a 'payment' to be transferred, conditional on the outcome observed. These contracts, and strategies for constructing them, will be discussed in more detail in Section 3.

As we are interested in structurally dynamic MASs, we may consider situations where the global population of agents is partitioned into smaller, temporary ad-hoc groups. We denote such a partitioning as $\mathcal{G} = \{G_1, \ldots, G_n\}, \forall G \in \mathcal{G}, G \subset A$. Members of $\mathcal{G}$ can then represent entities such as ad-hoc teams, coalitions or cliques.

2.2 Trust Evaluations

As our focus in this paper is on decision-making with trust evaluations, we do not discuss how these evaluations are formed here. Instead, we assume the existence of a probabilistic trust evaluation model (e.g. [Jøsang and Ismail, 2002; Teacy *et al.*, 2006]) which produces evaluations in the form of probabilities over task outcomes. This can be represented as a function $T^x_\tau : A \times 2^A \rightarrow \mathbb{R}$ which returns a real-valued trust evaluation, given a candidate, and a set of recommender agents.

We use $P^x_{y:\tau}(o)$ to denote the subjective probability, as perceived by a trustor x, of trustee y achieving outcome o in task τ. We refer to these evaluations as *unconditional* trust evaluations, as they are independent of the delegated agents effort choice. Similarly, $P^x_{y:\tau}(o|e)$ denotes the subjective probability, perceived by x, of a trustee y achieving outcome o in task τ, given that y selects the effort level e. Accordingly, trustors also maintain conditional trust models of the form $T^x_{e:\tau} : A \times 2^A \rightarrow \mathbb{R}$. We refer to these trust models, and their evaluations, as *conditional* models and evaluations, as they describe a trustors opinion of a trustee when a particular effort e is chosen.

Trustors build these models based on observations (or evidence) of the performance of trustees in different tasks (and effort levels, if trustors choose to monitor these). If a trustee has no direct evidence about a particular candidate, he can seek evidence from other agents in the society who may have interacted with that candidate before. However, as we consider structurally dynamic multi-agent systems, we cannot assume that the entire population of agents will be able to respond to delegation or reputation queries all of the time. When agents are formed in ad-hoc groups, communication (and, by extension, delegation and reputational evidence gathering) may only be possible within a trustors current group.

3 Delegation Strategies

Given that we have assumed a probabilistic trust evaluation model, we can take a decision-theoretic approach to the problem of trustworthy partner selection in situations involving risk. Figure 1 shows the problem from the trustors perspective, as a decision tree. Square nodes represent decision points for one of the agents, whereas circular nodes represent points where uncertainty is resolved. The leaves of the tree represent the final payoffs. The trustors goal is to determine the best action to take for the decision $D(x)$, i.e. whether to delegate, and if so, to whom. In the present model, trustors can choose between five delegation strategies:

1. *Simple delegation* ($Del(y, \tau, C^x_{y:\tau})$): the trustor delegates the task without considering the trustees effort choice, relying on the general (i.e. not specific to any effort level) trust evaluation (Section 3.1).
2. *Delegate with monitoring* ($Del_{Mon}(y, \tau, C^x_{y:\tau})$): the trustor invokes the trustee to adopt the desired effort level, and also pays the monitoring cost to observe which effort levels were selected by trustees. This permits the construction of *conditional* trust based on effort level choice (Section 3.3).
3. *Delegate without monitoring* ($Del_{NoMon}(y, \tau, C^x_{y:\tau})$): the trustor invokes the trustee to adopt the desired effort level, but does not monitor, and forfeits the ability to learn about the behaviours of trustees in different effort levels (Section 3.2).

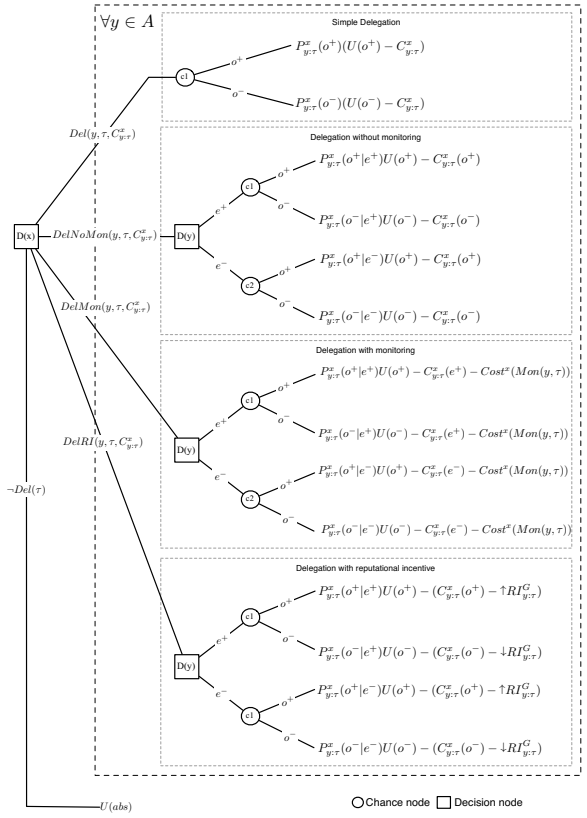

Figure 1: Decision tree for trustor agents.

4. *Delegate with reputational incentive* ($Del_{RI}(y, \tau, C^x_{y:\tau})$: trustors produce contracts which attempt to influence the trustees effort level choice by including the loss or gain the trustee would suffer from a change in reputation as an explicit incentive (Section 3.4).
5. *Abstain from delegation* ($\neg Del(\tau)$): the trustor abstains from delegation entirely if no agent can be found such that the benefits fail to outweigh the costs/risks of delegating, obtaining $U(abs)$.

Abstinence payoffs can represent the expected utility of delegating a less risky task, and obtaining a lower payoff. Trustors consider delegating riskier (and potentially more profitable) tasks first, falling back on less risky options if necessary.

The delegation process begins when a trustor x wishes to delegate some task τ. Firstly, the trustor observes the environment and obtains sets of potential candidates Y and recommenders R. Next, the trustor evaluates the candidates using his trust evaluation model T^x_τ, producing a set of opinions Ω^x_Y. For each candidate, the trustor computes the best contract from each of the alternative delegation strategies (i.e. unmonitored, monitored, etc.). The candidate y, whose best contract maximises the trustors EU, is selected. If the EU of agent y's contract is not greater than the abstention utility, then the trustor abstains. Otherwise, x proposes the contract to y, who can then accept or refuse. If y accepts, delegation proceeds. x observes and subjectively evaluates the outcome and may, by monitoring, observe the effort level choice of y. Based on these observations, the contract is enacted, transferring rewards and penalties between the agents as specified. If y refuses, x can seek another candidate from Y. Finally, the trust model T^x_τ is updated to reflect the new observation.

3.1 Simple Delegation

In the simplest case, the trustor makes no attempt to monitor the trustee. The cost of invoking the trustee is lost whether the delegation succeeds or not. The cost required to invoke the trustee to accept delegation is the cost of the desired level of trustee effort e, plus the utility the trustee could obtain by refusing the delegation (the trustees reservation utility), i.e.: $C^x_{y:\tau} = Cost^y(e^*) + V(abs)$. Both values are communicated by the trustee to the trustor, and need not be the same for every trustee under consideration.

The expected utility of delegating in this way to a particular agent y is given by $EU[Del(y, \tau, C^x_{y:\tau})] = \sum_{o \in \mathcal{O}_\tau} P^x_{y:\tau}(o)(U(o) - C^x_{y:\tau})$.

The problem with this kind of delegation, is that it permits the trustee full discretion in its choice of effort level. Furthermore, as the trustor has no knowledge about a candidates trustworthiness in any particular effort level, it is not possible to make an informed choice of effort e.

3.2 Delegation without Monitoring

Ideally, the trustor would have access to the effort-conditional trust evaluation $P^x_{y:\tau}(o_i|e_j)$. In this case, even though the trustees effort choice is hidden, the trustor can produce contracts which are conditional on the outcome that is observed.

The contract in this case is conditional on the observed delegation *outcome*. Given such a contract $C^x_{y:\tau}$, the trustees expected utility, from the point of view of the trustor, is then given by: $EV(e, C^x_{y:\tau}) = \sum_{o \in \mathcal{O}_\tau} P^x_{y:\tau}(o|e)V(C^x_{y:\tau}(o) - Cost^y(e^*))$

Through the contract, the trustor aims to influence the effort choice of the trustee, even though that choice will never be observed.

That is, the trustee's expected utility of providing an effort level e is the expected utility of the lottery over outcomes of τ, given the outcome-based contract $C^x_{y:\tau}$, minus the cost of expending the effort. The process of identifying a contract for a given candidate y in task τ then involves two stages [Grossman and Hart, 1983]:

1. For each possible effort level $e \in \mathcal{E}_\tau$, the trustor identifies the lowest cost contract $C^x_{y:e}$ that will invoke y to select effort e while still providing y with its minimum expected utility $V_y(abs)$.
2. Then, the trustor selects an effort level e which maximises his expected utility, given the costs of inducing that effort, and the conditional distribution $P^x_{y:\tau}(o_i|e_j)$.

The first step can be performed by solving, for all $e' \in \mathcal{E}_\tau$, the program $C^x_{y:e'} = \text{minimise}_{C^x_{y:\tau}} EV(e', C^x_{y:\tau})$, subject to the following constraints:

- *Incentive compatibility*: an ideal contract should align the interests of the trustee and trustor; the rational choice

for the trustee should be the choice of the trustor [Myerson, 1979], such that $e' = \arg\max_{e \in \mathcal{E}_\tau} EV(e, C^x_{y:\tau})$.

- *Participation constraint*: the contract to induce e' must give the trustee at least his reservation utility, such that $EV(e', C^x_{y:\tau}) \geq V(abs)$.

Once the trustor has obtained the least-cost contract for each effort level, the effort level e (and corresponding contract $C^x_{y:e*}$) which maximises the trustors expected utility will be chosen, given the cost of inducing that effort, by solving $e^* = \text{maximise}_{e \in \mathcal{E}_\tau} \sum_{o \in \mathcal{O}_\tau} P^x_{y:\tau}(o|e^*)(U(o) - C^x_{y:e}(o))$.

The best contract for y is then the contract which sustains the best level of effort, i.e. $C^x_{y:\tau} = C^x_{y:e*}$. The trustor's final utility function is then given by $EU[Del_{NoMon}] = \sum_{o \in \mathcal{E}_\tau} P(o|e^*)(U(o) - C^x_{y:\tau}(o))$.

3.3 Delegation with Monitoring

If a candidate cannot be trusted to select the desired effort level, the trustor can opt to monitor the candidates choice. Since the trustor monitors the trustees choice (and incurs monitoring cost $Cost^x(Mon^x(y, \tau))$), the contract in this case is dependant on the observed *effort level*. We then define $Cost^x(y, e)$ as the minimum incentive required to invoke y to expend effort e, as $Cost^x(y, e) = V_y(abs) + Cost^y(e)$.

The benefits of monitoring are unlikely to be without cost to the trustor. For example, a manager concerned about untrustworthy staff may be able to install CCTV cameras relatively cheaply; however, it may still be necessary to expend time and effort in checking the footage for indiscretions. Trustors must therefore be able to determine the expected value of a monitoring activity (EVM). A trustor computes the EVM for a candidate in the following way. Firstly, the expected utility of delegating to y, given a particular effort level e, is calculated as $EU[Del_{Mon}(y, \tau, e, C^x_{y:\tau})] = P^x_{y:\tau}(o^+|e)U(o^+) - Cost^x(Mon^x_{y:\tau}) - C^x_{y:\tau}(e)$.

Notice that the contract structure $C^x_{y:\tau}$ now maps the trustee's observed effort choice to payoffs. Next, we calculate the maximum effect a monitoring action can have on the trustor's expected utility in a future interaction. To do this, we use our trust evaluation model to simulate the observation of one positive and negative experience, and obtain new trust evaluations $P^x_{y:\tau}(o^+|e, o^+)$ and $P^x_{y:\tau}(o^+|e, o^-)$ representing the new trust ratings given the observation of a positive and negative outcome, respectively.

By substituting the real $P^x_{y:\tau}$ values for these speculative ones and re-evaluating the decision tree (Figure 1), we estimate the pair of expected utilities EU^+ and EU^- which we would obtain (in a subsequent interaction), given that the trustee has chosen effort level e. To compute the final value for EVM, we compare the absolute maximum difference between the EU of this decision without the results of monitoring, to the EU with the results, such that $EVM(Mon^x_{y:\tau}, e) = |EU - \max(EU^+, EU^-)|$. If the cost of monitoring is greater than its expected value, then it does not provide sufficient benefit to justify its use.

3.4 Delegating with Reputational Incentives

As we have mentioned, the potential losses or gains associated with reputational changes can also provide an incentive for trustees to select a particular effort level. We base our model on the notion that untrustworthy agents can expect to be selected for interaction with lower probability than trustworthy ones, and so must lower their reserve costs, with respect to trustworthy agents, in order to remain competitive.

In this way, less trustworthy trustees still have a chance to be selected in low risk interactions and repair their damaged reputation. However, in doing this, trustees obtain less utility from interactions, while having to expend the same utility performing the task. In this way, we can quantify the reputational incentive in a given interaction as the estimated utility loss a trustee will incur in its subsequent interaction as a result of poor performance in the current interaction.

We cannot expect trustees to know the success and failure payoffs for trustors, as these may vary between contexts and trustors. Therefore, we assume trustees compete with regards to the minimum *expected loss* (EL), or risk, that delegating to them entails. Given that trustees must obtain at least $V(abs)$, we can define a trustees EL in terms of its reputation and reserve. The EL of a trustee as perceived by a group of agents $G \subseteq A$ is then given by $EL^G_{y:\tau} = (1 - P^G_{y:\tau}(o^+))V_y(abs)$, where $P^G_{y:\tau}(o^+)$ denotes the probability of y achieving o^+ in task τ, as estimated by the group G (i.e. the *reputation* of y in task τ within group G).

By rearranging, we can find that the maximum reserve $V^*_y(abs)$ that y can expect in order to present a competitive EL to G is given by $V^*_y(abs) = \frac{EL^G_{y^*:\tau}}{1 - P(o^+)}$

Note that the utility a trustee can insist on while remaining competitive now depends on the trustees reputation within the group G. Both trustor and trustee can compute new prospective reputation values $\uparrow P^G_{y:\tau}(o^+)$ and $\downarrow P^G_{y:\tau}(o^+)$ in the event of a positive or negative outcome respectively by adding 1 (or more, if repeated unmonitored delegation is planned) 'simulated' observation to the aggregated evidence used to form the trust evaluation. Now we can define reputational incentives for both success and failure outcomes as the gain and reduction in maximum utility expectation in the subsequent interaction with trustors in G:

$$\uparrow RI^G_{y:\tau} = \frac{EL^G_{y^*:\tau}}{1 - \uparrow P^G_{y:\tau}(o^+)} - V^*_y(abs) \quad (1)$$

$$\downarrow RI^G_{y:\tau} = V^*_y(abs) - \frac{EL^G_{y^*:\tau}}{1 - \downarrow P^G_{y:\tau}(o^+)} \quad (2)$$

Delegating with reputational incentives proceeds similarly to unmonitored delegation. A solution is computed for each candidate, as in Section 3.2. This time, reputational incentives are also computed, and deducted from the contract, as they do not involve a real transfer of utility from trustor to trustee. Instead, reputational incentives can be communicated by x to y in the same way as explicit ones. Trustees can verify reputational incentives before delegation by estimating $\uparrow P^G_{y:\tau}(o^+)$ and $\downarrow P^G_{y:\tau}(o^+)$ from their own interaction histories in G. Alternatively, trustees can maintain their own trust evaluation models, and use these to evaluate their own reputation from the perspective of the society (G). A special case occurs when $G = \{x, y\}$, as y will always be y, and so

ID	Mean(e^-)	St.Dev(e^-)	Mean (e^+)	St. Dev (e^+)
p_1	0.6	0.15	0.9	0.05
p_2	0.4	0.15	0.6	0.15
p_3	0.3	0.10	0.4	0.15
p_4	0.0	1.00	0.3	0.10
p_5	0.0	1.00	0.0	1.00

Table 1: Trustee behaviours in effort levels e^- and e^+

reputational incentives available to x will always be 0. This is because y has an effective monopoly; trustors have no alternatives to choose from.

4 Evaluation

In order to evaluate our approach, we implemented the described model within a simulated multi-agent environment. We show that, in certain circumstances, decision-making and delegation using a mixture of trust and control can be beneficial, even when those controls are costly to implement. Our hypotheses in this section are as follows:

- *Hypothesis 1*: Trustors will perform better when using control strategies, as well as trust, than when using trust alone.
- *Hypothesis 2*: Agents will initially prefer to build trust by monitoring. As trust builds, unmonitored delegation will be preferred.

We now discuss the experimental methodology we employ to investigate these hypotheses.

4.1 Experimental Setup

Initially, 500 agents are created to play the role of trustees, and 40 agents to play the role of trustors. We create 20 ad-hoc groups within the society, each comprising 10 agents. The life time of each group is 5 interaction steps, after which it is disbanded, and a new group created in its place. In each interaction step, each trustor agent interacts with a trustee with an interaction probability $P_{interact} = 0.8$. We control the basic rate of dynamicity in the society with a join/leave probability parameter $P_{jl} = 0.01$, which determines the probability with which, in each interaction step, any agent (trustor or trustee) will leave the society, to be immediately replaced by a new agent from the same profile. Each experiment lasts for 500 interaction steps.

Trustees are drawn from a number of hidden profiles which determine their behavioural characteristics (Table 1). Profiles are associated with two gaussian distributions, representing the effort levels e^- and e^+. When an agent 'performs' a task, selecting a particular effort, the outcome is drawn from the corresponding gaussian distribution. Trustors consider outcomes above 0.5 to signify task success, and otherwise failure. The profile $p1$ represents a reliable class of agents, while $p4$ represents agents who will usually perform poorly. Profiles $p2$, $p3$ and $p4$ represent unreliable agents who may perform well or poorly, and agents of type $p5$ behave randomly in either effort level.

All trustors delegate the same task type τ. In our experiments, we define $U(o^+) = 12$, $U(o^-) = -1$ and $U(abs) = 0$ for all trustors. We set $Cost(Mon) = 2$. For all trustees, $Cost(e^+) = 2$, $Cost(e^-) = 1$. All trustees initially begin with $V(abs) = 1$.

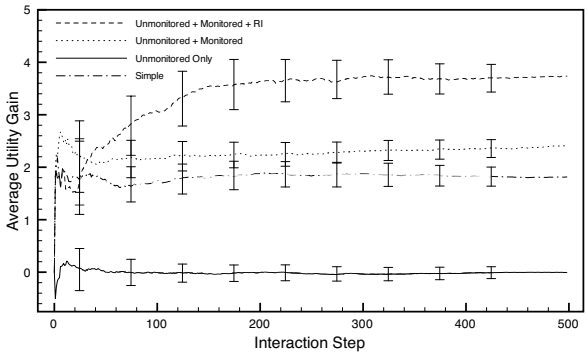

Figure 2: Comparison of society performance when different decision strategies are available.

- *Simple*: only unconditional trust evaluations are used.
- *Unmonitored*: trustee choice is considered, using conditional trust, but monitoring is not allowed.
- *Monitored*: both unmonitored and monitored strategies are available.
- *Reputational Incentive*: unmonitored, monitored and reputational incentive strategies are available.

In our simulations, our agents use Subjective Logic [Jøsang *et al.*, 2006] as their trust evaluation mechanism (thus implementing the T^x_τ function from Section 2.2). This does not affect the generality of our approach, which requires only that *probabilistic* estimates be produced.

4.2 Results

Hypothesis 1

Figure 2 shows the performance of our model in each of the four cases above, measuring the average change in utility experienced by the trustors in the society in each interaction. Error bars represent one standard deviation in the data. The *Simple* strategy obtained an average utility gain of 1.8 over the course of the experiment. In the *Unmonitored* case, the model performs very poorly, obtaining an average utility just above or below 0. This is to be expected, as this strategy has no way of obtaining conditional evidence, and so cannot build conditional trust. In the *Monitored* mode, the model performs much better, rising to around 2.4 by 500 interactions. The *Reputational Incentive* mode outperforms the others, achieving around 3.6.

A one-way ANOVA was conducted to compare the effects of the model on trustor performance. The results in these cases were found to be statistically significant using ANOVA testing at the $p < 0.001$ level [F(2,1494) = 11784, $p = 2.2 \cdot 10^{-16}$]. Post-hoc comparison using a Tukey HSD test indicated that the means in all cases were significantly different from each other. This supports our hypothesis that monitoring and reputational incentives can increase the decision-making performance.

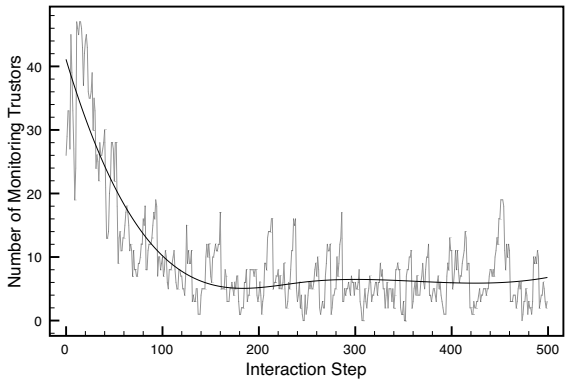

Figure 3: Decreasing monitoring activities over time.

Hypothesis 2
Figure 3 shows the number of trustors monitoring in each interaction step. Monitoring is initially preferred, as agents learn about the conditional behaviours of their partners. Over time, the informational benefit of monitoring decreases, and so agents begin to favour unmonitored delegation forms, using reputational incentive. Monitoring can never be completely avoided, however, as the population of agents is continually changing.

5 Discussion and Conclusions

Reputation mechanisms in multi-agent systems can play two distinct roles. Firstly, they enable trustors to make better evaluations by making use of the experiences accumulated by peers, and secondly, they provide an incentive for good behaviour among a societys trustees. In [Castelfranchi *et al.*, 2006], the authors discuss reputation as a form of *capital* that benefits trustees, as being trusted within a society increases the chance of being selected as an interaction partner, and increases the minimum 'price' a trustee can obtain in its interactions. In this paper, we have shown that this kind of reputation can play a role in supporting decisions in multi-agent societies.

We have shown that trust alone is necessary but not sufficient when making delegation decisions in the presence of uncertainty. In the presence of risk and uncertainty, trust is required to facilitate interactions, but interactions are required to obtain evidence from which trust can be built. In highly dynamic multi-agent systems, building trust is even more difficult. Approaches which aim to overcome the problem of building trust in dynamic environments [Burnett *et al.*, 2010; Hermoso *et al.*, 2010] still require a base of interactions from which to form generalisations. We have shown that, by employing controls in addition to trust, trustors can mitigate some of the perceived risk in their interactions, and be motivated to delegate, providing crucial initial interactions required to bootstrap trust.

We have made few assumptions about information assumed to be mutually shared between trustors and trustees. Specifically, we assume only that trustees effort and reserve costs are shared. Our approach can complement existing probabilistic trust evaluation approaches in order to support decision-making in multi-agent systems characterised by high dynamicity, uncertainty and risk.

Acknowledgements

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[Burnett *et al.*, 2010] C. Burnett, T. J. Norman, and K. Sycara. Bootstrapping trust evaluations through stereotypes. In *Proceedings. of 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 241–248, 2010.

[Castelfranchi *et al.*, 2006] Cristiano Castelfranchi, Rino Falcone, and Francesca Marzo. Being trusted in a social network: Trust as relational capital. *Trust Management*, pages 19–32, 2006.

[Grossman and Hart, 1983] S.J. Grossman and O.D. Hart. An analysis of the principal-agent problem. *Econometrica: Journal of the Econometric Society*, 51(1):7–45, 1983.

[Hermoso *et al.*, 2010] R. Hermoso, H. Billhardt, and S. Ossowski. Role Evolution in Open Multi-Agent Systems as an Information Source for Trust. In *Proceedings. of 9th International Conference on Autonomous Agents and Multiagent Systems*, 2010.

[Huynh *et al.*, 2006] T. D. Huynh, N. R. Jennings, and N. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Auton. Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[Jøsang and Ismail, 2002] A. Jøsang and R. Ismail. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.

[Jøsang *et al.*, 2006] A. Jøsang, R. Hayward, and S. Pope. Trust network analysis with subjective logic. In *Proceedings of the 29th Australasian Computer Science Conference-Volume 48*, pages 85–94. Australian Computer Society, Inc. Darlinghurst, Australia, Australia, 2006.

[Myerson, 1979] R.B. Myerson. Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, 47(1):61–73, 1979.

[Sabater and Sierra, 2005] J. Sabater and C. Sierra. Review on Computational Trust and Reputation Models. *Artificial Intelligence Review*, 24(1):33–60, 2005.

[Teacy *et al.*, 2006] W. Teacy, J. Patel, N. R. Jennings, and M. Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.