

Trust in agent-based software

Sarvapali D. Ramchurn and Nicholas R. Jennings
University of Southampton

While the Office of Science and Technology commissioned this review, the views are those of the authors, are independent of Government and do not constitute Government policy.

1 INTRODUCTION

Many modern computer applications are open distributed systems in which the (very many) constituent components are spread throughout a network, in a decentralized control regime, and which are subject to constant change throughout the system's lifetime. Examples include the grid (Foster and Kesselman 1998), peer-to-peer computing (Gong 2002), the semantic web (Berners-Lee et al. 2001), web services (Cerami 2002), e-business (Deitel et al. 2001), m-commerce (Sadeh 2002; Vulkan 1999), autonomic computing (Kephart and Chess 2003) and pervasive computing environments (Schmeck et al. 2002). In all of these cases, however, there is a need to have autonomous components that act and interact in flexible ways in order to achieve their design objectives in uncertain and dynamic environments (Simon 1996). Given this, agent-based computing has been advocated as the natural computation model for such systems (Jennings 2001).

More specifically, open distributed systems can be modelled as open multi-agent systems that are composed of autonomous agents that interact with one another using particular mechanisms and protocols. In this respect, interactions form the core of multi-agent systems. Thus, perhaps not unsurprisingly, the agent research community has developed rich models of interactions including coordination (Jennings 1993; Durfee 1999), collaboration (Pynadath and Tambe 2002; Cohen and Levesque 1990) and negotiation (Rosenschein and Zlotkin 1994; Kraus 2001; Jennings et al. 2001). However, their application in large-scale open distributed systems presents a number of new challenges. First, the agents are likely to represent different stakeholders that all have their own aims and objectives. This means the most plausible design strategy for an agent is to maximize its individual utility (von Neuman and Morgenstern 1944).

Second, given that the system is open, agents can join and leave at any given time. This means that an agent could change its identity on re-entering and avoid punishment for any past wrongdoing. For example, an agent could sell low quality products, leave the system as soon as it gets paid (so avoiding retribution from buyers or authorities operating in the system) and then subsequently rejoin the system unscathed. Third, an open distributed system allows agents with different characteristics (for example, policies, abilities, roles) to enter the system and interact with each other. Given this, agents are likely to be faced with a number of possible interaction partners with varying properties. For example, several agents might offer the same type of web service, but with different efficiencies (for example, speed of execution) or degrees of effectiveness (for example, providing richer forms of output).

Fourth, an open distributed system allows agents to trade products or services (for example, through various forms of auctions or market mechanisms) and collaborate (for example, by forming coalitions or virtual organizations) in very many ways. Therefore, agent designers are faced with a choice of a number of potential interaction protocols that could help them achieve their design objectives. Moreover, the choice about which interaction protocol (or mechanism) to adopt is important since each protocol may enforce a different set of rules of encounter and each protocol may result in a different outcome for the agents involved (for example, auctions maximize the utility of only two agents, the auctioneer and the buyer, while voting maximizes the social welfare).

Specifically, we can characterize the key interaction problems in such contexts through the following questions:

- How do agent-based system designers decide how to engineer protocols (or mechanisms) for multi-agent encounters?
- How do agents decide whom to interact with?

Sarvapali Ramchurn and Nicholas Jennings

- How do agents decide when to interact with each other?

In formulating a protocol, it is intended that the sequence of moves of the agents and the allocation of resources brought about by applying the protocol, are made in such a way that they prevent agents from manipulating each other (for example, through lies or collusion) so as to satisfy their selfish interests. Therefore, having such protocols in place provides guarantees that should facilitate the choice of interaction partners at any given time. However, protocols may, at times, be subject to trade-offs among the rules they enforce in trying to achieve their objectives (Sandholm 1999) (for example, in voting, while some constraints force truth-telling, they may lead to intractability). In such cases, it is left to the agents to decide how, when and with whom to interact without any guarantees that the interaction will actually achieve the desired benefits. To make such decisions would require agents to be fully informed about their opponents, the environment and the issues at stake. Such information should enable agents to devise probabilities for particular events happening and allow them to act in a way that maximizes their expected utility (Savage 1954). Moreover, given such information, agents should be able to act strategically by calculating their best response given their opponents' possible moves during the course of the interaction (Binmore 1992).

However, both the system (enforcing the protocol) and the agents may have limited computational and storage capabilities that constrain their control over interactions. Moreover, the limited bandwidth and speed of communication channels limits the agents' sensing capabilities in real world applications. Thus, in practical contexts it is impossible to reach a state of perfect information about the environment and the interaction partners' properties, possible strategies and interests (Binmore 1992; Russell and Norvig 1995; Axelrod 1984). Agents are therefore necessarily faced with significant degrees of uncertainty in making decisions (that is, it can be difficult or impossible to devise probabilities for events happening). In such circumstances, agents have to trust each other in order to minimize the uncertainty associated with interactions in open distributed systems.¹

In more detail, trust has been defined in a number of ways in different domains (see Falcone et al. (2001) for a general description). However, we find the following definition most useful for our purposes:

Trust is a belief an agent has that the other party will do what it says it will (being honest and reliable) or reciprocate (being reciprocative for the common good of both), given an opportunity to defect to get higher payoffs (adapted from Dasgupta (1998)).

Broadly speaking, there are two main approaches to trust in multi-agent systems.² Firstly, to allow agents to trust each other, there is a need to endow them with the ability to reason about the reciprocative nature, reliability or honesty of their counterparts. This ability is captured through trust models. The latter aim to enable agents to calculate the amount of trust they can place in their interaction partners. A high degree of trust in an agent would mean it is likely to be chosen as an interaction partner and (possibly) a reciprocative strategy used towards it over multiple interactions. Conversely, a low degree of trust would result in it not being selected (if other, more trusted, interaction partners are available) or a non-reciprocative strategy being adopted against it over multiple interactions (if there is no better alternative). In this way, trust models aim to guide an agent's decision making in deciding on how, when and with whom to interact. However, in order to do this, trust models initially require agents to gather some knowledge about their counterparts' characteristics, which can be done in many different ways (for example, through inferences drawn from the outcomes of multiple direct interactions with these partners or through indirect information provided by others). The direct interaction case leads us to consider methods by which agents can learn or evolve better strategies to deal with honest and dishonest agents such that payoffs are maximized in the long run. The indirect interaction case requires agents to be able to develop methods to reliably acquire and reason about the information gathered from other agents.

While trust models pertain to the reasoning and information gathering ability of agents, the second main approach to trust concerns the design of protocols and mechanisms of interactions (that is, the rules of encounter). These interaction mechanisms need to be formulated to ensure that those involved can be sure they will gain some utility if they rightly deserve it (that is, a malicious agent cannot tamper with the correct payoff allocation of the mechanism). Thus, we expect agents to interact using a particular mechanism only if it can be trusted. For example, an English auction can be trusted (by the bidders) to some extent since it ensures that the auctioneer cannot tamper with the bids because these are publicly voiced. However, the same auction cannot be trusted (by the

Sarvapali Ramchurn and Nicholas Jennings

auctioneer) to elicit the bidders' true valuation of the auctioned goods (because the dominant strategy of this mechanism is to bid lower than one's true valuation of the goods and a little higher than the current asking price). This highlights the need for protocols that ensure that the participants will find no better option than telling the truth and will interact honestly with each other.

As can be seen, trust pervades multi-agent interactions at all levels. With respect to designing agents and open multi-agent systems we therefore conceptualize trust in the following ways:

- individual-level trust, whereby an agent believes its interaction partners are honest or reciprocal;
- system-level trust, whereby the actors in the system are forced to be trustworthy by the rules of encounter (that is, protocols and mechanisms) that regulate the system.

The above approaches can be seen as being complementary. Thus, while protocols aim to ensure the trustworthiness of agents at the system level, they cannot always achieve this objective without some loss in efficiency and, in such cases, trust models at the individual level are important in guiding an agent's decision making. Similarly, where trust models at the individual level cannot cope with the overwhelming uncertainty in the environment, system-level trust models, through certain mechanisms, aim to constrain the interaction and reduce this uncertainty.

Generally speaking, the various issues concerning trust at these two levels have been dealt with separately in the literature, each forming a different piece of the puzzle, without any consideration for how they all fit together. To rectify this position, this paper presents a critique of the work that has been carried out on trust in multi-agent systems. More specifically, we evaluate the most prominent trust models that have been presented and show how they all fit together at the individual and at the system level. From this, we develop a general classification of approaches to trust in multi-agent systems and outline the open challenges that need to be addressed in order to provide a comprehensive view of trust in computational systems.

The paper is organized as follows. Section 2 deals with models that fit into the individual aspect of trust. We describe and evaluate the trust models that have been devised using learning and evolutionary techniques, reputation and socio-cognitive concepts. Section 3 deals with system level trust where we illustrate different mechanisms that enforce certain properties of the interaction and hence the trustworthiness of agents involved. Section 4 concludes and outlines the key future lines of research.

2 INDIVIDUAL-LEVEL TRUST

Here we take the viewpoint of an agent situated in an open environment trying to choose the most reliable interaction partner from a pool of potential agents and deciding on the strategy to adopt with it (that is, the who, when and how of interactions mentioned in Section 1). As we have already indicated, there are a number of ways the agent can go about doing this. Firstly, it could interact with each of them and learn their behaviour over a number of encounters. Eventually, it should be able to select the most reliable or honest agents from the pool or devise an appropriate strategy to deal with the less (or more) reliable ones. In this case, the agent reasons about the outcome of these direct interactions with others. Secondly, the agent could ask other agents about their perceptions of the potential partners. If sufficient information is obtained and if this information can be trusted, the agent can reliably choose its interaction partners. In this case, the agent reasons about interactions that others have had with its potential partners (indirect interactions). Thirdly, the agent could characterize the known motivations of the other agents. This involves forming coherent beliefs about different characteristics of these agents and reasoning about these beliefs in order to decide how much trust should be put in them.

Given the above, we can classify trust models at the individual level as either learning (and evolution) based, reputation based or socio-cognitive based. While the learning and evolutionary models aim to endow agents with strategies that can cope with lying and non-reciprocal agents, reputation models enable agents to gather information in richer forms from their environment and make rational inferences from the information obtained about their counterparts. Socio-cognitive models adopt a rather higher level view of trust that takes the knowledge of motivations of other agents for granted and proposes ways to reason about these motivations. The remainder of this section follows this classification and outlines the main models in each of the various categories.

2.1 Learning and Evolving Trust

In this section we consider trust as an emergent property of direct interactions between self-interested agents. Here we assume that the agents will interact many times rather than through one-shot interactions. This tallies with the concept of trust as a social phenomenon that is inherently based on multiple interactions between two parties (Molm et al. 2000; Carley 1991; Prietula 2000; Yamagishi et al. 1998; Dasgupta 1998). It is further assumed that agents have an incentive to defect (Dasgupta 1998). For example, defecting in an interaction could mean that the agent does not satisfy the terms of a contract, sells poor quality goods, delivers late or does not pay the requested amount of money to a seller. In these examples, defection could get higher payoffs for the agent defecting (for example, the seller gets paid more than the actual value of the goods sold) and cause some utility loss to the other party (for example, the buyer loses utility in buying a low quality product at a high price). However, defection reduces the possibility of future interactions since the losing agent would typically attempt to avoid risking future utility losses. In contrast, if both interaction participants cooperate, we assume that they get an overall higher payoff in the long run (Axelrod 1984). For example, a seller delivering goods on time or selling goods of an initially agreed quality may result in future purchases from the buyer. In all these cases, we are generally assuming that the agents already know the payoffs associated with each of their actions.

If the payoffs of each encounter are known, the agents can reason strategically by assessing the best possible move of their opponent and hence plan their own best response. This analysis falls within the realm of game theory (von Neuman and Morgenstern 1944), which regards all interactions as games with different payoffs (for example, winning or losing the game) for the individual players (that is, the interaction partners). Most games assume that the move of an opponent is not known in advance. In such one-shot games, the safest (that is, minimizing possible loss), and not necessarily the most profitable, move will be chosen unless there can be some way to ascertain that the other party can be trusted.³ Thus, if an agent believes its counterpart is reciprocative then the former will never defect, otherwise it will, and both could end up with lower payoffs than if they trusted each other or learned to trust each other. This belief may only be acquired if the game is repeated a number of times such that there is an opportunity for the agents to learn their opponent's strategy or adapt to each other's strategy.

To this end we will first consider models that show how trust, through reciprocation (of positive deeds), can be learned or evolved over multiple direct interactions (Section 2.1.1). These interaction models, however, greatly simplify the interactions to extreme notions of cooperation and defection. In reality we believe these two extremes can rather be considered the two ends of an axis measuring the success of the outcome of the interaction. In this context, cooperation could mean, for example, that a seller actually delivers some of the goods (rather than not delivering all), but some slight delay in the delivery might still be considered poor cooperation (rather than complete defection). Hence, the perception of an agent of another party's trustworthiness is relative to the level of satisfaction of the outcome. We therefore consider, in Section 2.1.2, how the payoffs in the individual interactions can actually be modelled in realistic applications.

2.1.1 Evolving and learning strategies

The most common example used to illustrate the evolution of trust or cooperation over multiple interactions is Axelrod's tournaments revolving around the prisoner's dilemma (Axelrod 1984).⁴ Within very controlled settings, Axelrod's tournaments have shown that the tit-for-tat strategy was the most successful (reaping higher average points over all the encounters) relative to other selfish or nicer (that is, mostly cooperative) strategies. Tit-for-tat cooperates on the first move and imitates the opponent's move in the remaining interactions. By adopting this strategy, agents are, in fact, trusting each other, but would punish untrustworthy behaviour should it ever happen (and also forgive if trustworthy behaviour were shown again). If two agents adopt tit-for-tat (or permanently cooperative strategies) it is shown that they end up with the highest payoffs compared to all other strategies. However, when faced with other selfish strategies, tit-for-tat does not get the maximum payoff, though it actually gets a higher payoff than most other strategies. This is because tit-for-tat loses on the first encounter and actually never beats an opponent.

It is therefore required that an agent adapts its strategy according to the type of environment (agents therein) it encounters in order to minimize losses and foster cooperation. By allowing agents to adapt, Wu and Sun have shown that trust can actually emerge between them (Wu and

Sarvapali Ramchurn and Nicholas Jennings

Sun 2001). This means that the agents evolve a trusting relationship (that is, a cooperative strategy) by evaluating the benefit of each possible strategy over multiple interactions. A multi-agent bidding context, in which a number of seller agents bid for contracts in an electronic marketplace, is chosen to exemplify the concept. It is first shown that when agents are all nice (always cooperating) to each other, sellers tend to learn to exploit them. To counter this, the nice agents learn to use tit-for-tat to minimize their losses. As a result, the nasty sellers (exploitative agents) then learn to be reciprocative since cooperating would bring them more benefit than defecting in the long run. Thus, trust emerges as a result of the evolution of strategies over multiple interactions. This example also shows that the evolution of strategies allows nice agents to beat nasty ones in the long run. However, while strictly applying to the bidding context, Wu and Sun's model does not take into account the fact that there might be some utility loss (in the short run) in cooperating with the other party (for example, giving away some resources).

In this respect, while acknowledging a cost to cooperation, Sen⁵ demonstrates how reciprocity can emerge when the agents learn to predict that they will receive future benefits if they cooperate (Sen 1996). In a more recent set of experiments, Sen and Dutta (2002) give clear guidelines about evolutionary stable strategies (not necessarily tit-for-tat) in different types of environments (with different sorts of strategies). They show that collaborative liars (collaborating defectors) perform well whenever the number of interactions is small and the number of philanthropic agents (always cooperating) is large. However, reciprocative strategies performed better in all other scenarios they tested. Besides proving that reciprocation pays, these results show that the length and number of interactions matter when it comes to evaluating another agent's trustworthiness. If the number of interactions is too low, then trust cannot be built. This is corroborated by Mui et al. (2002) in their probabilistic trust model which identifies a threshold for the number of encounters needed to achieve a reliable measure of an opponent's trustworthiness based on performance appraisal.

In the case that this threshold cannot be reached, other techniques must be used to elicit trustworthiness. In this respect, Mukherjee et al. (2001) have shown how trust can be acquired if agents know their opponent's chosen move in advance. They show that, in the case where agents do not reveal or only partially reveal (only the first mover does) their actions before their opponent acts, no amount of trust is built since it is optimal for the opponent to always choose to defect. However, in the bilateral information revealing scenario (both agents reveal their actions), both agents trust each other through mutually learning to choose an action that results in higher outcomes than predicted for the non-learning situation. It is to be noted that their model (as well as Sen's), besides assuming a static environment, uses an arbitrarily defined function to calculate the cost of interacting and returns from future actions (the basis of which might need more investigation, but has proven to be quite successful in applications simulated). Also, the experiments carried out have shown conclusive results only with a particular pay-off matrix. Hence, more experiments would be needed to show the applicability of the results to the more general class of interactions with different payoff matrices.

Up to this point, all the models deal strictly with the problem of cooperation between self-interested parties. However, not all multi-agent interactions are strictly competitive. For example, agents may be self-interested, but still may need to achieve a maximum pay-off as a group or society since the latter determines their individual payoffs (for example, individuals contributing an unspecified amount of money to build a road in their community – the total amount collected decides whether the road will be built, giving utility to the individuals, otherwise the money is used for a secondary purpose). This is the problem tackled by Birk (2000, 2001). It is thus shown that trust may not only emerge from the evolution of strategies (Birk 2000), but can also arise strictly out of learning (Birk 2001). The learning method Birk exposes uses a continuous case N-prisoner's dilemma as basis for simulation. This involves agents contributing to a common fund required for the society to achieve its goals, but each agent is tempted to contribute less than the equal split of the total investment required in the hope that others will contribute more. In this context, a cooperative strategy (that is, contributing more than the equal split) gradually predominates in an environment where bad agents (that is, contributing less) are in the majority. This is because the low investment obtained by the society impacts negatively on the utility of each individual member as well, forcing these latter to learn to cooperate to get higher pay-offs. However, as the number of cooperative agents increases, the agents learn to defect again to get better pay-offs (this is similar to what Wu and Sun's model predicts). Birk's results additionally show that the society reaches an equilibrium with a high level of trust (or cooperation) among its members.

The above learning and evolutionary models of multi-agent strategic interactions assume complete

Sarvapali Ramchurn and Nicholas Jennings

information (for example, strategies, payoff matrix) for the multi-agent learning algorithms to work. These results have typically been obtained through simulations using very strict assumptions and settings rather than real life scenarios where the main assumption of complete information about payoffs simply does not hold. Also, most of the learning models conceive the outcome of interactions as being bi-stable, that is, either a defection or cooperation. To be more realistic, we believe agents need to infer, from the information gathered through their direct interactions, how their opponents are performing and how their performance is affecting their goals. This leads on to devising realistic trust metrics.

2.1.2 Trust metrics

For an agent to computationally model its trust in its opponent, it is first required that the former can ascribe a rating to the level of performance of its opponent. The latter's performance over multiple interactions can then be assessed to check how good and consistent it is at doing what it says it will. Therefore, in addition to a performance rating, an agent also needs a means of keeping track of the performance of an agent (in its direct interactions with it). To this end, Witkowski et al.'s (2001) model proposes a trust model whereby the trust in an agent is calculated based on its performance in past interactions (the context is a trading scenario for an intelligent telecommunication network where bandwidth is traded, the quality and quantity of which is varied depending on the trust suppliers and buyers have in each other). The update to the trust value is different for the different types of agents defined in the system. Specifically, consumers update their trust value according to the difference between their bids and the received goods (bandwidth in this case). The better the quality (size) of the goods the higher the increase in trust, and conversely for low quality goods. A higher trust in a seller would then result in it being chosen for future purchases (conversely for low trust). In contrast, the supplier agents update their trust in the consumers according to the extent to which the quality (size) of the goods (bandwidth) supplied has been exploited. If the quality offered was not fully used, then the trust goes down since it implies that the consumer has dishonestly asked for more than is actually needed. If the quality is fully exploited, the trust goes up. Results of the experiments show how trust (of consumers in suppliers) is effectively strongly dependent on the ability of suppliers to cope with the demand.

The model used by Witkowski et al. simplifies the calculation of trust through equations that deal with measurable quantities of bandwidth allocation and bandwidth use. Other models (such as Mui et al. 2002; Sen and Sajja 2002; Schillo et al. 2000) consider the performance of an agent to be simply a bi-stable value (good or bad). While these models achieve the objectives of the agents for the specific simulation settings studied, they cannot generally be used more widely because realistic interactions in an open distributed system involve richer outcomes (for example, quality of goods traded, efficiency of task handling, duration of task). To overcome this, we need more generic means of assessing performance over time. To this end, Sabater and Sierra (2002) do not just limit the overall performance to a bi-stable value or to an efficiency measure (as per Witkowski et al. 2001), and rather attribute some fuzziness to the notion of performance. Thus, depending on the context, the performance of an agent can be subjectively judged on a given scale where -1 represents very poor performance, 0 represents neutral, and +1 represents being very good. The REGRET system actually gives richer semantics to ratings (or impressions) by defining their particular characteristics. For example, an agent can express a satisfaction -0.5 for the delivery date of some goods and +1 for the price of the same goods. These impressions are then analyzed and aggregated using fuzzy reasoning techniques to elicit a representative value for the overall impression (or trust) of one agent on another.

In contrast to Witkowski et al.'s model, REGRET's evaluation of trust is not only based on an agent's direct perception of its opponent's reliability, but it also evaluates its behaviour with other agents in the system. This is carried out because only perceiving direct interactions can pose a number of problems. For example, in an open system, it would be very difficult for an autonomous agent to select an interaction partner if the agent itself had never interacted with another party (that is, it has no history to analyze). Moreover, the method opens itself to attack by strategic liars that knowing how they are rated by the other side, can adapt their behaviour (for example, clients overloading their channels) to make the other party believe they are trustworthy (that is, fully using their bandwidth). In such cases an agent could be better off evaluating other environmental parameters (such as asking other agents about their impressions of each other) in an attempt to get a more reliable rating of its opponents. However, a number of problems arise in doing this. For example, information gathered from other agents could be wrong or incomplete. Such problems

2.2 Reputation Models

Reputation can be defined as the opinion or view of someone about something (Sabater and Sierra 2002). Here we consider that this view can be mainly derived from an aggregation of opinions of members of the community about one of them.⁶ In multi-agent systems, reputation can be useful when there are a large number of agents interacting (for example, online auctions, stock-trading). Reputation should, for example, enable buyers to choose the best sellers in the system. Moreover, reputation can induce sellers to behave well if they know they are going to be avoided by future buyers as a result of their reputation going down due to bad behaviour. These different aspects of reputation divide the field into the following lines of research:

- devising methods to gather ratings that define the trustworthiness of an agent, using relationships existing between members of the community;
- devising reliable reasoning methods to gather as much information from the aggregation of ratings retrieved from the community;
- devising mechanisms to promote ratings that truly describe the trustworthiness of an agent.

The last of the above items is dealt with in Section 3.2 (since it falls within the realm of system-level trust). For now we will be concerned with the first two items because these are at the level of individual agents.

In order to organize the retrieval and aggregation of ratings from other agents most reputation models borrow from sociology the concept of a social network (Burt 1982; Buskens 1998). Similar to human societies, this assumes that agents are related to each other whenever they have roles that interconnect them or whenever they have communication links (for example, by observation, direct communication or as information sources) established between one another. Through this network of social relationships, it is assumed that agents, acting as witnesses to interactions, can transmit information about each other (Panzarasa et al. 2001). Information takes the form of a performance rating (for example, good or bad, seller delivers late, buyer never paid) as explained in Section 2.1.2. Such a rating could then be shared by the different nodes of the social network, thus giving rise to the concept of reputation.

2.2.1 Retrieving ratings from the social network

Yu and Singh (2002a) tackle the problem of retrieving ratings from a social network through the use of referrals. In this context, referrals are pointers to other sources of information similar to links that a search engine would plough through to obtain a web page or URL. Through referrals, an agent can provide another agent with alternative sources of information about a potential interaction partner (particularly if the former cannot handle the latter's request itself). Yu and Singh propose a method of representing a social network (based on a referral network (Singh et al. 2001)) and then provide techniques to gather information through the network (Yu and Singh 2003). Specifically, they show how agents can explore a network by contacting their neighbours and can use referrals gathered from the latter to gradually build up a model of the social network. Furthermore, Schillo et al. (2000) enrich the representation of an existing social network by annotating nodes of the network to represent their particular characteristics. Thus, each node of the network holds two values: (i) the trust value, which describes the degree of honesty of the agent represented by the node, and (ii) the degree of altruism (that is, being good to others even at the expense of one's own utility). Both of these values are used to deduce the trustworthiness of witnesses queried at the time of calculating the reputation of potential interaction partners (see Section 2.2.2). From an established social network it is then possible to derive higher level concepts. For example, Sabater and Sierra (2002) and Yu and Singh (2002a) derive the concept of a group of neighbours from the social network by identifying those nodes (agents) that are close together (linked together). Thus, having a social network represented allows an agent to select and contact those agents it needs in order to get a proper measure of the reputation of another agent. For example, Yu and Singh's model takes into account ratings from those agents that are close (by virtue of the number of links separating them with a potential interaction partner) to choose witnesses for a particular agent. Underlying this is the assumption that closer witnesses will return more reliable ratings.

It is further assumed in all of the above models, that witnesses share ratings freely (that is, without

Sarvapali Ramchurn and Nicholas Jennings

any profit). This is a relatively strong assumption, which can be removed if proper mechanisms are implemented (as will be seen in Section 3.2). Therefore, given that agents have represented their social network and properly extracted the ratings of their counterparts from the network, they then need to aggregate these ratings so as to form a coherent impression of their potential interaction partners.

2.2.2 Aggregating ratings

Several means of aggregating ratings in online communities already exist. For example, in eBay (2003), ratings are +1 or -1 values (in addition to textual information) that are summed up to give an overall rating. Such simplistic aggregation of ratings can be unreliable, particularly when some buyers do not return ratings (see Kollock (1999) and Resnick and Zeckhauser (2002) for a complete account of online reputation systems). For example, a sum of ratings is biased positively when there are fewer people not reporting bad ratings even though these people have had bad experiences. Having no rating is considered neither as a bad rating, nor as a good rating and is simply discarded from the aggregation. Moreover, ratings are open to manipulation by sellers trying to build their reputation. While the latter problem can be dealt with by designing sophisticated reputation mechanisms (see Section 3.2), the former problem can be solved at the level of the agent's reasoning mechanism.

To this end, Yu and Singh (2002b) deal with absence of information in their reputation model. The main contribution of their work is in aggregating information obtained from referrals while coping with the lack of information. More specifically, they use the Dempster Shafer theory of evidence to model information retrieved (Yager et al. 1994). The context is the following: an agent may receive good or bad ratings (+1 or -1) about another agent. When an agent receives no rating (good or bad), how should it classify this case of no information? In Yu and Singh's model, a lack of belief (or disbelief) can only be considered as a state of uncertainty (where all beliefs have an equal probability of being true). Dempster's rule allows the combination of beliefs obtained from various sources (saying an agent is trustworthy, untrustworthy, or unknown to be trustworthy or not) to be combined so as to support the evidence that a particular agent is trustworthy or not. Moreover, together with a belief derived from ratings obtained, an agent may hold a belief locally about the trustworthiness of another due to its direct interaction with it. However, in such cases, the ratings obtained from witnesses are neglected. Nevertheless, their measure of reputation does not discredit nor does it give credit unnecessarily to agents (as eBay does) in the absence of information.

As can be seen, Yu and Singh do not deal with the possibility that an agent may lie about its rating of another agent. They assume all witnesses are totally trustworthy. However, an agent could find some benefit in lying about its rating of an opponent if it is able to discredit others such that it appears to be more reliable than them. In this respect, Schillo et al. (2000) deal with the problem of lying witnesses. They first decompose the rating into social metrics of trust and altruism (as discussed above – see Section 2.2.1). The latter metrics are used in a recursive aggregation over the network taking into consideration the probability that the witnesses queried may lie to (or betray) the querying agent. In this way, the value obtained for the trust in an agent is more reliable than fully trusting witnesses as in the case of Yu and Singh's model (which assumes cooperative settings). The probability of a witness lying to the querying agent is actually learnt over multiple interactions in Schillo et al.'s model.

Similarly, Sen et al. (2000) extend this work and demonstrate how agents can cope with lying witnesses in their environment through learning rather than attributing subjective probabilities to the event of a witness lying (Sen and Sajja 2002). Specifically, they develop a reputation model which makes the same simplifying assumptions as those illustrated in Section 2.1. Their approach shows how the sharing of trust values (or reputation) can benefit reciprocative agents in the long run. In the short run though, selfish and lying agents still benefit from totally reciprocative agents. Furthermore, it is shown that, over time, colluding agents cannot exploit reciprocative agents if these learn the behaviour of the former and share their experience with others of a similar type. The reciprocative agents then become selfish towards these lying and completely selfish agents so as to minimize utility loss in interacting with them. Their model, however, fails when the number of witnesses in the environment falls below a given threshold. This is because a sufficiently high number of witnesses is needed to report ratings about most lying agents in a population. If this is not the case, there is a higher probability of a reciprocative agent interacting with a lying one that has not been encountered by the witnesses previously.

Sarvapali Ramchurn and Nicholas Jennings

While Yu and Singh's model demonstrates the power of referrals and the effectiveness of Dempster Shafter's theory of evidence in modelling reputation, Schillo et al.'s, and Sen et al.'s models show how witness information can be reliably used to reason effectively against lying agents. These models, however, greatly simplify direct interactions and fail to frame such interactions within the social setting (that is, relative to the type of relationships that exists between the witnesses and the potential interaction partners). To overcome this limitation, Sabater and Sierra (2002) adopt a (sociological) approach closer to real life settings. Thus, the reputation value, which is representative of the trust to be placed in the opponent, is a weighted sum of subjective impressions derived from direct interactions (the individual dimension of reputation), the group impression of the opponent, the group impression of the opponent's group and the agent's impression of the opponent's group (together, all of these compose the social dimension of reputation). Now, the weights on each term allow the agent to variably adjust the importance given to ratings obtained in these diverse ways. Moreover, older ratings, devised as shown in Section 2.2.1, are given less importance relative to new ones.

The strong realism of REGRET also lies in its definition of an ontological dimension that agents can share to understand each other's ratings (for example, a travel agent being good might imply low price for one agent, but for another might imply good quality seats reserved). However, REGRET does not handle the problem of lying (strategically) among agents. Ratings are obtained in a cooperative manner (from an altruistic group) rather than in a competitive setting (where witnesses are selfish). Moreover, the aggregation method REGRET uses can be sensitive to noise since ratings are simply summed up. In contrast, Mui et al.'s (2002) model calculates the probability of an agent being trustworthy on the next interaction by considering the frequency of (positive and negative) direct impressions conditional upon the impressions gathered from the social network. This approach, we believe is less sensitive to noisy ratings from the network.

2.3 Socio-cognitive Models of Trust

The approaches to modelling trust at the individual level that we have considered in the previous sections are all based on an assessment of the outcomes of interactions. For example, learning models consider the pay-offs of each individual strategy, while reputation models assess outcomes of both direct and indirect interactions (that is, third-party assessments). However, in assessing the trustworthiness of an opponent, it may also be important to consider the subjective perception of the latter since it enables a more comprehensive analysis of the characteristics of the opponent (Dasgupta 1998; Gambetta 1998). For example, the tools and abilities available to the other party could be assessed to check whether or not the agent can indeed use these to carry out an agreed task. Such beliefs or notions are normally stored in an agent's mental state and are essential in assessing an agent's reliability for doing what it says it will (that is, being capable), or its willingness to do what it says it will (that is, being honest).

In this respect, we report the line of work initiated by Castelfranchi and Falcone (1998, 2000b,a; Falcone and Castelfranchi 2001). In particular, they highlight the importance of a cognitive view of trust (particularly for Belief-Desire-Intention agents (Wooldridge 2002)) in contrast to a mere quantitative view of trust (Sections 2.1 and 2.2). The context they choose is that of task delegation where agent *x* wishes to delegate a task to agent *y*. In so doing agent *x* needs to evaluate the trust it can place in *y* by considering the different beliefs it has about the motivations of agent *y*. They claim the following beliefs are essential (in *x*'s mental state) to determine the amount of trust to be put in agent *y* by agent *x* (these have been adapted and summarized):

- competence belief: a positive evaluation of *y* by *x* saying that *y* is capable of carrying out the delegated task as expected. If agent *y* is not capable, there is no point in trusting it to accomplish the task fully;
- willingness⁷ belief: *x* believes that *y* has decided and intends to do what it has proposed to do. If agent *y* is not believed to be willing to do the task, it might be lying if it says it wants to do so. This would then decrease *x*'s trust in *y*;
- persistence⁷ belief: *x* believes that *y* is stable enough about its intention to do what it has proposed to do. If *y* is known to be unstable, then there is added risk in interacting with *y*, hence a low trust would be put in *y* even though it might be willing to do the task at the point the task is delegated;
- motivation belief: *x* believes that *y* has some motives to help *x*, and that these motives will probably prevail over other motives negative to *x* in case of conflict. This highlights the

Sarvapali Ramchurn and Nicholas Jennings

possibility for y to defect as argued in Section 2.1. The motives mentioned here are the same as the long term gains obtainable in helping x achieve its goals. If y is believed to be motivated (to be helpful or positively reciprocative as in Section 2.1), then x will tend to trust it.

To devise the level of trust agent x can place in agent y , agent x would need to consider each of the above beliefs (and possibly others). These beliefs actually impact on trust, each in a different way, which need to be taken into account in a comprehensive evaluation of all beliefs concerned. For example, the competence belief is a prerequisite for trusting another agent, while the motivation belief would vary according to the calculation of the future payoffs to the agents over multiple interactions. This kind of strategic consideration becomes even more important when such beliefs are known to all actors (that is, the preferences of agents are public). For example, what could happen if agent y knows that x trusts it, or relies on it? The authors claim that this may increase the trustworthiness of x in y 's mind, the self-confidence of y , or its willingness to serve x , which in turn change the trustworthiness of y . Agent x can then take into account the possible effects of its trust in y (even before performing the delegation) to support its decision of delegating. However, Castelfranchi and Falcone's approach is strongly based on humans who are not always rational beings (as opposed to what we expect agents to be).

In contrast to the cognitive approach of Castelfranchi and Falcone, Brainov and Sandholm (1999) support the need to model an opponent's trust (as described above) with a rational approach (they specifically target the context of non-enforceable contracts). They do so by showing that if an agent has a precise estimation of its opponent's trust (in the former), this leads to maximum payoffs and trade between the two agents. However, if trust is not properly estimated, it leads to an inefficient allocation of resources between the agents involved (hence a loss in utility) since both under estimate or over estimate their offers on exchanged contracts. It is also shown that it is in the best interests of the agents, given some reasonable assumptions, to actually reveal their trustworthiness in their interaction partner (to efficiently allocate resources)!

Although still in its infancy, the socio-cognitive approach to modelling trust takes a high level view of the subject. However, it lacks the logical grounding (as shown by Brainov and Sandholm) in rational mechanisms which learning and reputation models (and mechanisms) provide. In effect, the socio-cognitive approach could exploit the assessment performed by these models to form the core beliefs illustrated above. Thus, speaking generally, all the individual models of trust could contribute to a comprehensive evaluation of trust at the individual level. This would take into account strategies learned over multiple interactions, the reputation of potential interaction partners, and finally the latter's believed motivations and abilities regarding the interaction. However, it can be computationally expensive for an agent to reason about all the different factors affecting its trust in its opponents. Moreover, as highlighted earlier, agents are limited in their capacity to gather information from the various sources that populate their environment. Given these constraints, instead of imposing the need to devise trust at the individual level, it can be more appropriate to shift the focus to the rules of encounter so that these ensure that interaction partners are forced to be trustworthy. In this way, these rules of encounter can, at times, compensate for limited applicability of individual-level trust models (conversely, whenever the rules of encounter cannot guarantee interacting agents will be trustworthy, we might need to resort to individual-level trust models to do this).

3 SYSTEM-LEVEL TRUST

In the context of open multi-agent systems, we conceive of agents interacting via a number of mechanisms or protocols that dictate the rules of encounter. Examples of such mechanisms include auctions, voting, contract-nets, market mechanisms and bargaining, to name but a few. These mechanisms take agents to be completely self interested and, therefore, need to make sure that the rules of encounter prevent lying and collusion between participants. Generally speaking, such requirements impose some rigidity on the system (for example, an English auction forces bidders to reveal their bids). However, these rules enable an agent to trust other agents by virtue of the different constraints imposed by the system. These constraints can be applied in a number of ways. Firstly, it is possible to engineer the protocol of interaction such that the participating agents find no gain in utility from lying or colluding. Secondly, an agent's reputation as being a liar (or truthful) can be spread by the system. Thus, knowing that their future interactions will be compromised if they are reputed to be liars, agents can be forced to act well, up to the point they

Sarvapali Ramchurn and Nicholas Jennings

leave a system. Thirdly, agents can be screened upon entering the system by requiring proof of their reliability through the references of a trusted third party.

Against this background, we subdivide system-level trust in terms of (i) devising truth-eliciting interaction protocols, (ii) developing reputation mechanisms that foster trustworthy behaviour, and finally (iii) developing security mechanisms that ensure new entrants can be trusted. This is the structure that we adopt in the following subsections.

3.1 Truth-eliciting Interaction Protocols

In order to ensure truth-telling on the part of agents involved in an interaction, a number of protocols and mechanisms have been devised in recent years (see Sandholm (1999) for an overview). These protocols aim to prevent agents from lying or speculating while interacting (for example, lying about the quality of goods sold or proposing a higher price than one's true valuation for goods to be bought). They do so by imposing rules dictating the individual steps in the interaction and the information revealed by the agents during the interaction. Thus, by adhering to such protocols it is expected that agents should find no better option than telling the truth. Given the aim of this paper, we do not wish to launch into a detailed explanation of all available protocols that have such properties and enforce them to a certain degree (see Sandholm (1999) for such an analysis). Rather we will focus on one such protocol (namely auctions, since these are the most widely used mechanism in multi-agent system applications).

As opposed to other mechanisms, such as voting which focuses on the social welfare, auctions focus on the benefit of only two agents – the auctioneer and the bidder. There are four main types of single-sided auctions, namely the English, Dutch, first-price-sealed-bid, and Vickrey. In the English auction, each bidder is free to raise his bid until no bidder is willing to raise any further, thus ending the auction. The Dutch auction instead starts with a very high asking price and reduces it in steps until one of the bidders bids for the item and wins the auction. The first-price-sealed-bid involves agents submitting their bids without knowing others' bids. The highest bidder wins the auction. In the Vickrey auction, the bids are sealed but the winner pays the price of the second highest bid.

In this context, the Dutch and English auctions enforce truth-telling on the part of the auctioneer (for example, the winner and the winning price cannot be faked) since bids are made publicly as opposed to Vickrey and first-price-sealed-bid auctions where the bids are hidden. However, the Dutch, English, and first-price-sealed-bid auctions do not ensure that the bidders reveal their true valuation of the goods at stake. This is because the dominant strategy in these auctions is to reveal either a lower valuation (in the case of Dutch and first-price-sealed-bid) or to bid only a smaller amount more than the asking price up to one's true valuation (in the case of the English auction). In contrast, the Vickrey auction does enforce truth-telling by bidders. A bidder's dominant strategy is to bid its true valuation since doing otherwise, given uncertainty about other bids and the final price to be paid, would result in some loss in utility. Bidding higher than its true valuation could end up with the agent paying more than its valuation and bidding lower than its true valuation could make it lose the auction altogether.

As pointed out above, the main weakness of the Vickrey mechanism is that it does not ensure truth-telling on the part of the auctioneer. The latter could still lie about the winning bid since bids are private and known only to the auctioneer (and obviously to each of the bidders in private, unless there is some amount of collusion). The auctioneer could thus ask for a higher price than the second highest bid (just below the highest bid) to the highest bidder. In so doing, the auctioneer reaps a higher benefit than it should, without the bidders knowing. In this respect, Hsu and Soo (2002) have implemented a secure (that is, ensuring the privacy of bids and the allocation of the goods to the true winner) multi-agent Vickrey auction scheme. The scheme differs from the original Vickrey auction in that it involves an additional step of choosing the auctioneer from among the bidders (advertised on a blackboard). The bidders submit their encrypted bids to a blackboard. The auctioneer is selected at random from the bidders and given a key to access all sealed bids and, using this key, it can only compare the bids' values. Thus, the auctioneer can only determine the order of bids and allocate the second highest bid to the winner. This scheme also allows the auctioneer (also a bidder), the winner, and the second highest bidder to verify the result by using their key to check the bids shown on the blackboard.

However, the Vickrey auction and the other main ones stated above are not collusion proof. This

Sarvapali Ramchurn and Nicholas Jennings

means that agents can collaborate to cheat the mechanism by sharing information about their bids. Collusion would first necessitate that the agents know each other before they place their bids and therefore arrange to place bids that do not reveal their true preferences (for example, agents withholding their bids in a Dutch auction until the asking price has gone very low, or some bidders colluding with the auctioneer to artificially raise the asking price in an English auction to force others to pay a very high price or bidders colluding to beat competitors in a Vickrey auction). To prevent this collusion from happening, Brandt (2001, 2002) extends the work of Hsu and Soo by devising a collusion proof auction mechanism that ensures the privacy and correctness of any $(M+1)$ st-price auction (that is, an auction where the highest M bidders win and pay a uniform price determined by the $(M+1)$ st price). In this type of auction, bids are sealed and the highest bid wins the auction, but pays a price determined by the auctioneer (for example, in the Vickrey auction the second highest price is paid).

Only the auctioneer and the bidder know the highest bid. To allow bidders to verify whether the winning bid is actually the highest (hence checking the trustworthiness of the winner and auctioneer) the protocol devised by Brandt distributes the calculation of the selling price between the individual buyers using some cryptographic techniques. However, the only other agent, apart from the seller, able to calculate the exact value of the selling price is the winner of the auction. The protocol also ensures that bids are binding. These conditions, combined with the fact that the protocol can be publicly verified, allow the identification of malicious bidders that would have tampered with the bids and prevent collusion from affecting a single bidder.

While being very powerful, the protocol is computationally expensive when a large number of agents is involved, but works well for small numbers.

As can be seen, most auctions are not robust to lying and collusion unless some security mechanism is plugged into them (that is, using cryptographic techniques). The protocols mentioned above, besides constraining interactions, neglect the fact that the agents in an open distributed system might want to interact more than one time. As was shown in Section 2.1, reciprocative or trustworthy behaviour can be elicited if agents can be punished in future interactions or strictly prevented from engaging in future interactions if they do not interact honestly. For example, if a winning bidder in an auction has been found to have lied about its preferences, it could be prevented from accessing future runs of the auction (Brandt 2002). If an agent knows it will lose utility in the future due to bad behaviour in the present, it will find no better option but to act in a trustworthy way. In this respect, earlier in the paper (see Section 2.1) we showed how agents could learn to actually adapt their strategy (reciprocative or not) in order to maximize their long term pay-offs against different strategies over multiple runs of an auction.

However, as pointed out in Section 1, open multi-agent systems allow agents to interact with any other agent in the environment. This could permit malicious agents to move from group to group whenever they are detected by a given group of agents and therefore exploit trustworthy agents as they move around. In order to prevent this from happening, agents can be made to share their ratings of their opponent with other agents in the environment once they have interacted with them. Techniques to allow agents to gather ratings and aggregate those in a sensible way were presented in Section 2.2. However, it was shown that these techniques do not consider the fact that we expect agents to share (true) ratings only if it brings them some utility. In open multi-agent systems, this can be achieved through reputation mechanisms which we discuss in the next section.

3.2 Reputation Mechanisms

As was seen in Section 2.2, the reputation models described do not take into account the fact that the agents are selfish and, therefore, will not share information unless some benefit can be derived from doing so. Furthermore, these reputation models (for example, REGRET or Yu and Singh's model) do not motivate the use of reputation by some agents to elicit good behaviour from other agents. These models aim to endow agents with a better perception of their opponent and do not consider the effect of doing so on an opponent when the latter is aware of it! Given these shortcomings of reputation models, reputation mechanisms consider the problem of inducing trustworthy behaviour and modelling the reputation of agents at the system level. Reputation mechanisms can operate through centralized or distributed entities that store ratings provided by agents about their interaction partners and then publicize these ratings, such that all agents in the environment have access to them. In this case, it is the system that manages the aggregation and

Sarvapali Ramchurn and Nicholas Jennings

retrieval of ratings as opposed to reputation models, which leave the task to the agents themselves. In so doing, reputation mechanisms can be used to deter lying and bad behaviour on the part of the agents. Moreover, reputation mechanisms aim to induce truthful ratings from witnesses and actually make it rational for agents to give ratings about each other to the system. Such a mechanism, that makes it rational for participants to use it, is said to be incentive compatible (Resnick and Zeckhauser 2002).

More specifically, Zacharia and Maes (2000) have outlined the *desiderata* for reputation mechanisms particularly with regard to how ratings are aggregated and how these impact on the behaviour of the actors in the system. They do not propose such requirements for agent-based reputation systems per se, but as we move into agent-mediated electronic commerce (He et al. 2003), it is obvious that such mechanisms will guide agent-based reputation systems. These *desiderata* are listed below:

1. It should be costly to change identities in the community. This should prevent agents from entering the system, behaving badly, and coming out of the system without any loss of utility or future punishment bearing upon them.
2. New entrants should not be penalized by initially having low reputation values attributed to them. If new entrants have low reputation they are less favoured though they might be totally trustworthy. This actually makes the system less appealing to agents (with bad reputation) intending to (re)enter the system.
3. Agents with low ratings should be allowed to build up reputation similar to a new entrant. This allows an agent to correct its behaviour if it has been shown to be badly behaved in the past.
4. The overhead of performing fake transactions should be high. This prevents agents from building their own reputation.
5. Agents having a high reputation should have a higher bearing than others on the reputation values they attribute to an agent. This presupposes that agents with high reputation will give truthful ratings to others. However, this can be contentious if reputation determines the level of profit the agent acquires since it could lead to the creation of monopolies or cartels in the market.
6. Agents should be able to provide personalized evaluations. This involves giving more than just a simple rating of +1 to -1 to allow a better evaluation of the reputation of another agent. For example, the REGRET system actually implements richer ratings that can be shared using the ontological dimension.
7. Agents should keep a memory of reputation values and give more importance to the latest ones obtained. This is needed to keep the reputation measure as up to date as possible and helps prevent an agent from building up positive reputation by interacting well and then starting defecting (the last defection having a greater effect than its past good behaviour).

With respect to the above requirements, Zacharia and Maes present two reputation systems (for example, for chatrooms, auctions and newsletters): SPORAS and HISTOS. While these are not strictly multi-agent systems, they present techniques to aggregate ratings intelligently and reflect the real performance of users in an online community. In both cases, the aggregation method allows newer ratings to count more than older ones. SPORAS, however, gives new entrants low initial reputation values and, therefore, reduces their chance of being selected as possible interaction partners. This is a trade-off afforded to prevent identity switching because an agent having low reputation would not be any better off by re-entering the system with a new identity. HISTOS is an enhancement to SPORAS, which takes into account the group dynamics as in REGRET. In particular, HISTOS looks at the links between users to deduce personalized reputation values (that is, taking into account the social network). This enables an agent to assemble ratings from those it trusts already rather than those it does not know. Moreover, both HISTOS and SPORAS have been shown to be robust to collusion. This is because those agents that are badly rated themselves have a diminished effect on the reputation of others and those they might want to protect. However, as the authors themselves point out, the major drawback is that users are reluctant to give bad ratings to their trading partners. This is because there is no incentive to give ratings in the first place (that is, it is not incentive compatible).

In an attempt to make reputation mechanisms incentive compatible, Jurca and Faltings (2003) introduce side payments to make it rational for agents to share reputation information. Thus, agents can buy and sell reports to and from special information agents supplied by the system. Reports are values between 0 and 1, where 0 represents completely bad behaviour and 1

Sarvapali Ramchurn and Nicholas Jennings

represents absolutely trustworthy behaviour. Agents are only allowed to sell a report for an agent when they have previously bought reputation information for that agent. This ensures that agents cannot sell reputation information they make up by themselves. They additionally propose two conditions to make a reputation mechanism robust to lying witnesses:

- agents that behave as good citizens (report truthfully) should not lose any utility;
- agents that give false reports should gradually lose utility.

In their model, they aim to fulfil the above conditions by implementing information agents which pay only for reports (to one agent) if they match the next report given by another agent (having interacted with the same agent as the previous one). In this way, the authors claim that agents revealing truthfully get paid and those that do not will lose money in buying reports and not getting paid on selling them. However, this method does not work if most agents lie about the reports or if they collude over giving false reports. Moreover, they assume that information agents already store some reputation information after bootstrapping the system. This overly simplifies the process of reputation management and, additionally, does not take into account the case of new entrants into the system. More work is therefore needed to make this model applicable to open multi-agent systems.

Jurca and Faltings's reputation mechanism actually aims to be generic and, as a result, suffers the above shortcomings. It might be preferable instead to design reputation mechanisms that are tailored to individual protocols of interaction. In this respect, Dellarocas (2002) introduced 'Goodwill Hunting' (GWH) as a more realistic feedback mechanism, for a trading environment. This system claims to:

- induce sellers of variable quality goods to truthfully reveal the quality of their goods;
- provide incentives to buyers to truthfully reveal their feedback.

The GWH algorithm uses the threat of biased future reporting of quality (of goods to be sold) in order to induce sellers to truthfully declare the individual qualities of their items. Specifically, the mechanism keeps track of the seller's 'goodwill'. This value represents the seller's trustworthiness. It is adjusted by the quality reported by buyers. Good reports bias goodwill positively and bad reports bias it negatively. To induce sellers to reveal the true quality of their goods, the goodwill factor is used to adjust the quality they wish to broadcast for the goods they wish to sell. Thus, if the seller has low goodwill, the quality of the goods it tries to publicize will be actually shown to have a lower quality by the system.

To induce buyers to report their ratings of sellers, they are given rebates on future transactions in the system. It is then shown that, if buyers report untruthfully, they can drive out sellers of good quality goods and, therefore, lose the opportunity of buying high quality goods. However, the mechanism makes several somewhat unrealistic assumptions about real life online markets. For example, it assumes that sellers are monopolists; that is, they are the only ones to sell a particular product (of varying quality). Also it assumes that buyers will interact with sellers only once. These assumptions are needed to simplify the analysis of the model. As Dellarocas points out, among other enhancements, it is still to be shown how the mechanism fares against strategic reporting from buyers whereby they force a seller to reduce the price of its goods by giving it bad ratings, hence damaging its reputation.

The reputation mechanisms detailed above and interaction mechanisms discussed in Section 3.1 actually try to enforce trustworthy behaviour by minimizing the opportunity for agents to defect to gain higher payoffs (see our definition of trust in Section 1. As has been shown, more of these mechanisms still need to be developed. In the case where interaction protocols and reputation mechanisms cannot guarantee trustworthy behaviour, there still exists a need to give agents in an open system the possibility of proving their trustworthiness and to enable other agents to recognize them as reliable interaction partners. One way this could proceed is by providing references from highly recognized sources. This is similar to the case of a job seeker providing credentials to a potential new employer. Note that this process is not the same as reputation building and acquisition, which pertains to the recognition of an entire community. Rather, credential assessment falls within the realm of network security, which we discuss next.

3.3 Security Mechanisms

In the domain of network security,⁸ trust is used to describe the fact that a user can prove it is who it says it is (Mass and Shehory 2001). This normally entails that it can be authenticated by trusted third parties (that is, those that can be relied upon to be trustworthy and as such are authorities in the system (Grandison and Sloman 2000)). At a first glance, this does not completely fit with our initial definition of trust, but it is certainly a basic requirement for the trust models and mechanisms described earlier to work (see Sections 2.1, 2.2, 2.3, 3.1, 3.2). This is because these models are based on the fact that agents can be recognized by their identity and would, therefore, require authentication protocols to be implemented.

To this end, Poslad et al. (2002) have recently proposed a number of security requirements that they claim are essential for agents to trust each other and each other's messages transmitted across the network linking them (that is, to ensure messages are not tampered with by malicious agents):

- identity: the ability to determine the identity of an entity. This may include the ability to determine the identity of the owner of an agent;
- access permissions: the ability to determine what access rights must be given to an agent in the system, based on the identity of the agent;
- content integrity: the ability to determine whether a piece of software, a message, or other information has been modified since it was dispatched by its originating source;
- content privacy: the ability to ensure that only the designated identities can examine a message or other data. To the others, the information is obscured.

The authors specify these requirements for the FIPA (Foundation for Intelligent Physical Agents) (2001) abstract architecture. These basic requirements can be implemented by a public key encryption and certificate infrastructure (Grandison and Sloman 2000). A digital certificate is issued by a certification authority, or CA, and verifies that a public key is owned by a particular entity. The public key in a certificate is also used to encrypt and sign a message in a way that only its owner can examine the content and be assured about its integrity. The two most popular public key models are PGP (Pretty Good Privacy) and the X.509 trust model (Adams and Farrel 1999). PGP supports a web of trust in that there is no centralized or hierarchical relationship between CAs, while the X.509 is a strictly hierarchical trust model for authentication (Grandison and Sloman 2000). However, these authenticating measures do not suffice for open multi-agent systems to ensure that agents act and interact honestly and reliably towards each other. They only represent a barrier against agents that are not allowed in the system or only permit their identification in the system. In order to enforce good behaviour in the system, it is instead possible that certificates are issued to agents if these agents meet specific standards that make them trustworthy.

In order to achieve this, trusted third parties are needed to issue certificates to agents that satisfy the standards of trustworthiness (that is, being reciprocative, reliable and honest). For example, agents would need to satisfy certain quality standards (for instance, products stamped with the Kitemark or the 'CE' marking are assured to conform to the British standards and the European community standards respectively) and terms and conditions for the products they sell (for instance, sellers have to abide by a 14-day full refund return policy in the UK for any goods they sell). It is only upon compliance with these quality standards that the agent would be able to sell its products. To this end, Herzberg et al. (2000) present a policy-based and certificate-based mechanism that can assign roles to new entrants. A certificate in this work is signed by some issuer and contains some claims about a subject. There is no restriction on what the claims can be. For example, they may be claims about organization memberships (company employee, etc.), the capabilities of the subject, or even the trustworthiness (or reliability) of the subject in the view of the issuer.

The mechanism in Herzberg et al. also enables a party to define policies for mapping new entrants to predefined business roles. Thus, an agent can ensure that a new entrant will act according to the settings defined by its role or access rights. The role assigned to an agent carries with it a number of duties and policies it needs to abide by. If the agent undertakes the role, it is forced to abide by the given rules of good behaviour. The process of role assignment and access provision is performed in a fully distributed manner, where any party or agent may be a certificate issuer. Moreover, it is not required that certificate issuers be known in advance. Instead, it is enough that,

Sarvapali Ramchurn and Nicholas Jennings

when requested, an agent that issues certificates provides sufficient certificates from other issuers to be considered a trusted authority according to the policy of the requesting party. This allows distributed trust build-up among parties in an open environment (Mass and Shehory 2001).

Mass and Shehory (2001) extend the work in Herzberg et al. (2000) to open multi-agent systems. Specifically, they take into account the fact that agents with reasoning or planning components can adapt their strategies rather than sticking to one strategy while maintaining their role (as discussed in Section 2.1). This means that an agent's role does not fully constrain its actions so as to prevent it from reasoning strategically about its interactions with other agents. An agent could thus learn how to adapt its strategy according to the role it has. For example, an agent bearing the role of accountant in a system could report fictitious profits, thus benefiting its company's share price, while still satisfying its role. To prevent such strategic defection or wrong doing, the agent assigning the role to the new entrant is allowed to adjust its priorities or policy based on results from interactions with others dynamically. This presents a more realistic view of using trust (both at the individual and system level) to decide how to constrain the actions (or strategies) of an interaction partner.

4 CONCLUSION AND FUTURE DIRECTIONS

This paper has systematically analyzed the issue of trust in open multi-agent systems. We have deliberately taken a broad-based approach in order to produce a comprehensive view of this multifaceted topic. In particular, we have related the different means of devising trust firstly at the individual level and then at the system level. These two approaches lay the burden of computation on the agent and the system, respectively. In effect, they complement each other by minimizing risks in different circumstances differently and aim to solve the same problem of deciding the 'who, when and how' of interactions.

At the individual level, we have described learning and evolutionary models that show how agents could evolve or learn more reciprocative strategies in order to get higher payoffs in the long run. Various means of characterizing the experience from individual interactions were presented and these were shown to lead to a measure of trust that enables an agent to choose future interaction partners and shape these interactions based on its personal experience with them. In contrast, reputation models have been shown to be efficient at gathering the experiences of others in various ways and using these to deduce the level of trustworthiness of another party. Various ways of gathering ratings from other parties using a social network were discussed. In so doing, we illustrated how the problem of lying witnesses can be dealt with using learning and probabilistic techniques and how agents can deal with the lack of ratings from the network. Having described the various ways of gathering information about direct and indirect experiences, it was then shown how an agent could use this information to form various beliefs about its counterparts. The socio-cognitive approach to trust also takes into account the fact that other beliefs about an agent's capabilities and motivations are essential in judging its trustworthiness.

Trust being enforced by the system was first discussed at the level of the interaction protocols and mechanisms themselves. We showed how the system can be devised so as to force the agents to be trustworthy. We particularly illustrated how auctions could be made secure and foster truth revelation on the part of bidders. We then showed how the threat of future punishment (through avoidance of or constraining interaction(s)) could be used by reputation mechanisms to prevent agents from lying about their preferences or force them to behave well in an open environment. Various methods of aggregating ratings and incentivizing agents to return ratings were discussed. The use of reputation through certificates was also shown to be an important solution in security mechanisms. The latter also ensure that agents are properly authenticated and therefore present a first line of defence against malicious agents in open multi-agent systems.

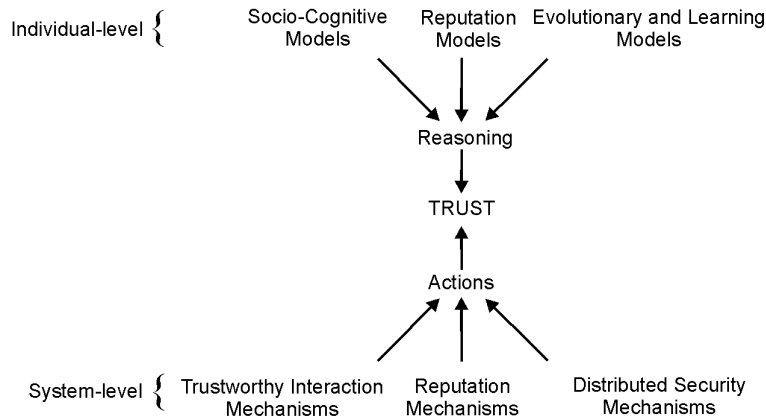
As can be seen from Figure 1, while the individual level trust models enable an agent to reason about its level of trust in its opponents, the system level mechanisms aim to ensure that these opponents' actions can actually be trusted. In more detail, using their trust models, agents can:

- reason about strategies to be used towards trustworthy and untrustworthy interaction partners (for example, being reciprocative or selfish towards them) given a calculation of pay-offs over future interactions;
- reason about the information gathered through various means (for example, either directly or through reputation models) about potential interaction partners;

Sarpapali Ramchurn and Nicholas Jennings

- reason about the motivations and capabilities of these interaction partners to decide whether to believe in their trustworthiness.

Figure 1 A classification of approaches to trust in multi-agent systems



In contrast, the mechanisms and protocols described (that is, enforcing system-level trust) aim to force agents to act and interact truthfully by:

- imposing conditions that would cause them to lose utility if they did not abide by them;
- using their reputation to promote their future interactions with other agents in the community or demote future interactions whenever they do not behave well;
- imposing specified standards of good conduct that they need to satisfy and maintain in order to be allowed in the system.

Based on our survey of the state-of-the-art, pertaining to both individual- and system-level trust in multi-agent systems, the key findings in this paper are as follows.

- Various trust models have been developed based on sociology, machine learning techniques and game theory. These have been shown to be useful in helping agents interact better. However, these models each look at a different facet of the trust problem without relating to each other (a lacuna which we fill in this paper).
- A very small number of interaction protocols have been shown to be trustworthy. The reason is that the computational complexity of interaction protocols can be a barrier to designing trustworthy interaction mechanisms.
- Security mechanisms provide a number of techniques to make interactions secure. However, they do not control the semantics of interactions beyond the line of defence provided by security policies and encryption techniques.
- Most trust models and interaction protocols do not cope effectively in the face of strategic lying. This means that agents can lie about their properties or the information they supply in order to exploit loopholes in the interaction protocol or in the decision making of their counterparts so as to gain more utility.
- Most trust models and interaction protocols are not collusion-proof. Agents can collude in order to exploit other agents or the system itself. Unless more attention is given to its detection and prevention, collusion could substantially alter the outcomes of interaction mechanisms and therefore damage open markets.
- While using game theory as a means of studying interactions, protocol designers and trust models make many unrealistic assumptions about the context comprising the environment and the social network. A more precise modelling of the context of interactions is needed. Trust models and interaction protocols should be then adapted to whichever (dynamic) context they are used in.
- With the advent of open distributed systems, agents representing different countries, institutions or societies will be interacting. This could give rise to a clash of norms and cultures (for example, laws, societal norms) which, in turn, will result in agents making the

Sarvapali Ramchurn and Nicholas Jennings

wrong assumptions about their counterparts, leading to distrust. Trust models should therefore conceptualize the difference in expectations arising from a difference in norms and cultures.

In the next sub-section, we outline future work that would be needed to completely define comprehensive trust models and mechanisms. Moreover, Appendix 6.1 provides a motivating scenario that shows how these abstract trust concepts can be grounded in a particular application context.

4.1 Research Challenges

We end our analysis of the state-of-the-art in trust in multi-agent systems by outlining the key issues that need to be solved in order to have a comprehensive trust model for open multi-agent systems.

Computational complexity – very few interaction protocols have been shown to elicit trustworthy behaviour from the interacting agents (for example, English auction, Vickrey auction). This is mostly because richer forms of interactions (for example, combinatorial auctions) can become intractable or computationally expensive. In order to deal with this problem, better protocols, with minimum computational complexity, need to be devised to regiment the interactions of agents in open societies.

Security mechanisms – security mechanisms provide a first line of defence against malicious attacks from outsiders. However, encryption and authentication tools do not control the semantics of direct interactions between agents in the system. To do so, security mechanisms would need to detect any unaccounted for behaviour in the system and, combined with trust models, direct agents to the most appropriate remedies.

Strategic lying – while some reputation mechanisms and models try to deal with this problem (for example, Schillo et al. (2000), Sen and Sajja (2002), Zacharia and Maes (2000)), most models do not give a deep treatment of strategic lying. Strategic lies aim to trick agents into believing the liars are trustworthy while allowing the liars to exploit these unaware agents. A more thorough treatment is needed to address this shortcoming both at the individual level and at the system level of trust.

Collusion detection – very few existing reputation or interaction mechanisms can prevent or deal with collusion (Sen and Sajja 2002; Brandt 2002). Moreover, while it has been shown how agents can learn to reciprocate good actions over time, it has not been shown how they could learn to collude, which is equivalent to reciprocating to only some agents and sharing false information about these accomplices to exploit others. We could expect agents to collude in an open environment and, if the system is to be robust and incentive compatible, collusion should be prevented. Otherwise, agents could end up trusting others that are exploiting them.

Context – most trust models do not take into account the fact that interactions take place within a particular organizational and environmental context (with the exception of the socio-cognitive approach to some extent). If an agent has performed poorly due to changes in its environment, it should not be taken to be dishonest or a liar. Rather, there should be the possibility to take into account the environmental variables in deciding to trust another agent. This necessitates a better evaluation of risks present in the environment (Molm et al. 2000; Yamagishi et al. 1998). If risks are high due to the lack of stringent rules of encounter (for example, preventing lying), an agent interacting honestly would then be considered to be more trustworthy than if the protocol of interaction dictated truth-telling for example. Thus, if rules prevent lying, there is no need to increase trust in interaction partners if they interact well since there is no guarantee they would still do so if the rules were not present (Molm et al. 2000).

Expectations – none of the models surveyed showed how agents could convey their expectations (about the outcome of interactions) to each other (for example, about the quality of goods exchanged or time of delivery). This we believe is important because, in an open environments, agents can have different concepts or ontologies that describe the expectations from an interaction. For example, 'high quality service' could mean 'timely delivery of goods' for one agent while the other party implied 'good price' in the former's ontology. REGRET presents such an ontological dimension of trust ratings that are shared, but does not show how this dimension could be shared between interaction partners to better understand each other's expectations about the outcomes of the interaction. Understanding these expectations would enable an agent to satisfy

Sarvapali Ramchurn and Nicholas Jennings

them in the way that they are understood to be from the other side. Otherwise an agent could be deemed untrustworthy because of its ignorance of the real expectations of another party.

Social networks – while in most reputation models or security mechanisms (to some extent) it is assumed that there exists a social network, the connections between the nodes in the network are rarely, if at all, given a meaning (that is, the semantics of connections are not detailed). Connections have mostly been used to represent past interactions among the agents in the community (that is, a connection means that an interaction has occurred between the two nodes at its ends) or are simply given to the agents (Sabater and Sierra 2002; Yu and Singh 2002b; Schillo et al. 2000). A clearer definition of relationships (for example, as collaborators, partnerships in coalitions, or members of the same organizations) defining the connections within the network would be needed. This, we foresee, should enable a better aggregation and evaluation of ratings, and hence trust.

APPENDIX 1

We choose the semantic web to illustrate the practical applications of trust for open multi-agent systems. This is because, while potential applications of agent-based systems such as ubiquitous computing and pervasive computing applications are still in their infancy, the semantic web is building upon the considerable success of the World Wide Web and technologies associated with it. Moreover, the semantic web is inherently strongly motivated by concepts in multi-agent systems (for example, reasoning under uncertainty, ontologies, communication languages). It can therefore be considered that the semantic web will provide the test-bed for the first large scale application of agent-based systems in everyday life. For these reasons, we provide the following vision of the semantic web (adapted from Berners-Lee et al. (2001)) and detail the roles of trust models and interaction mechanisms within it.

Lucy and Pete have to organize a series of appointments to take their mother to the doctor for a course of physical therapy sessions. (We identify the need for trust at each step of the scenario in italics.)

At the doctor's office, Lucy instructed her semantic web agent through her handheld web browser. The agent promptly retrieved information about Mom's prescribed treatment from the doctor's agent, looked up several lists of providers, and checked for the ones in-plan for Mom's insurance within a 20 mile radius of her home and with a rating of excellent or very good on trusted rating services.

The first interaction between Lucy's agent and the doctor's agent should involve a secure authentication protocol (see Section 3.3) that would ensure that Lucy's agent is allowed to handle her Mom's data. This protocol would first verify the true identity of Lucy's agent and assign to it the proper rights to handle the data. Also, the trusted rating services could be based on reputation mechanisms (see Section 3.2). These reputation mechanisms could publish the ratings of health care providers and reward agents which return ratings with discounts on treatment costs to be paid to the advertised providers. This would make the mechanism incentive-compatible. Also, different providers could bid, via a trusted mechanism such as a secure Vickrey auction, to provide the requested service to Lucy's agent (see Section 3.1). Provider agents would need to bid their true valuation of the treatment plan requested to win the bid whereas Lucy's agent would act as the auctioneer in this case.

Lucy's agent then began trying to find a match between available appointment times (supplied by the agents of individual providers through their web sites) and Pete's and Lucy's busy schedules. In a few minutes the agent presented them with a plan. Pete didn't like it: University Hospital was all the way across town from Mom's place, and he would be driving back in the middle of rush hour. He set his own agent to redo the search with stricter preferences about location and time. Lucy's agent, having complete trust in Pete's agent in the context of the present task, automatically assisted by supplying access certificates and shortcuts to the data it had already sorted through.

The interaction between individual providers and the user agents (Lucy's and Pete's) needs a secure mechanism that ensures messages transmitted between all parties are not manipulated. Pete's agent could enhance the search for trustworthy potential providers by looking at its past interaction history with them (see Section 2.1) rather than looking at only the reputed ones (see Sections 2.2 and 3.2). It could also use referrals of other agents in the network to get in touch with a trustworthy agent it does not directly know.

Almost instantly the new plan was presented: a much closer clinic and earlier times, but there were two warning notes. First, Pete would have to reschedule a couple of his less important

Sarvapali Ramchurn and Nicholas Jennings

appointments. He checked that they were not a problem. The other was something about the insurance company's list failing to include this provider under physical therapists: 'Service type and insurance plan status securely verified by other means' the agent reassured him. 'Details?'

Here the issue of reputation and distributed security is again raised (Sections 2.2 and 3.3). The 'other means' that helped to check the validity of the insurance company may pertain to an analysis of the certificates it provided that linked it to trusted sources. These certificates could provide evidence of the provider's compliance with laws and regulations of the country or certain quality standards that are equivalent to those needed by the insurance company.

Lucy registered her assent at about the same moment Pete was muttering, "Spare me the details," and it was all set. (Of course, Pete couldn't resist the details and later that night had his agent explain how it had found that provider even though it was not on the proper list.)

Here, the need for an agent to demonstrate how it could flexibly deal with different beliefs it acquired in the environment about potential interaction partners is highlighted (see Section 2.3). This implies a higher level reasoning ability than just an evaluation reputation of providers for example. The agent should also be able to reason about the selected provider's location and treatment facilities to decide on whether to trust that provider in being able to supply the required services.

NOTES

- 1 In agent-based systems, the network security issues associated with distributed systems (e.g. delegation, authorization) are not specifically dealt with. Some attempts are described briefly in Section 3.3 and we point to other relevant literature in the network security paradigm.
- 2 We here omit a third approach which is that taken in the mobile agent community. The latter focuses on identification issues (similar to network security approaches) and host-client interaction at the code-execution level rather than at the semantics of interactions which we are more concerned with in most agent-based systems.
- 3 The moves chosen will also be dependent on the risk attitude (risk seeking, risk neutral, or risk averse) of the agent. In this respect, we conceive of trust as a means to minimize the risk perceived by the agent (Yamagishi et al. 1998; Molm et al. 2000; Dasgupta 1998).
- 4 The prisoner's dilemma is a game involving two prisoners that have to decide whether to cooperate by not revealing their accomplice's deeds or to defect by revealing this information. The dilemma arises as a result of each other having to separately (in different rooms) decide to cooperate or not, resulting in some years of imprisonment (5 for one cooperating and one defecting, 3 for both if they both defect and 1 for both if they both cooperate). In the face of such uncertainty the dominant strategy proves to be defection even though this does not lead to best outcomes (hence the dilemma).
- 5 For a wider reading on the problem of learning cooperative strategies in competitive settings, see Mukherjee et al. (2001), Biswas et al. (2000) and Sen (1996).
- 6 We here distinguish between trust and reputation in the sense that the former is derived from direct interactions while reputation is mainly acquired (by an agent about another) from the environment or other agents and ultimately leads to trust. This distinction is only made to facilitate the study of the different models presented rather than to prescribe such an approach to trust and reputation.
- 7 In order to have this belief, agent x needs to model the mental attitudes of agent y.
- 8 We do not wish to give a complete account of network security mechanisms since this is beyond the scope of this paper. Rather, we will focus on the main concepts and models that strictly pertain to multi-agent systems. For a wider reading on network security for open distributed systems, see Grandison and Sloman (2000).

REFERENCES

- Adams, C. and Farrel, S. (1999), 'RFC2510 Internet x.509 Public Key Infrastructure Certificate Management Protocols. Technical Report', *The Internet Society*,
<http://www.cis.ohio-state.edu/htbin/rfc/rfc2510.html>, accessed 12 Apr. 04.
- Axelrod, R. (1984), *The Evolution of Cooperation*, New York: Basic Books.
- Berners-Lee, T., Hendler, J. and Lassila, O. (2001), 'The Semantic Web', *Scientific American* 284(5): 34-43.
- Binmore, K. (1992), *Fun and Games: A Text on Game Theory*, Boston MA: D. C. Heath and Company.
- Birk, A. (2000), 'Boosting Cooperation by Evolving Trust', *Applied Artificial Intelligence* 14(8): 769-84.
- Birk, A. (2001), 'Learning to Trust' in R. Falcone, M. P. Singh and Y. Tan (eds), *Trust in Cyber-societies, Integrating the Human and Artificial Perspectives, Lecture Notes in Computer Science, Vol. 2246*, Berlin: Springer, pp. 133-44.
- Biswas, A., Sen, S. and Debnath, S. (2000), 'Limiting Deception in Groups of Social Agents', *Applied Artificial Intelligence Journal* (Special issue on Deception, Fraud and Trust in Agent

Sarvapali Ramchurn and Nicholas Jennings

Societies) 14(8): 785-97.

Brainov, S. and Sandholm, T. (1999), 'Contracting with Uncertain level of Trust' in *Proceedings of the 1st ACM Conference on Electronic Commerce*, New York: ACM Press Press, pp. 15-21.

Brandt, F. (2001), 'Cryptographic Protocols for Secure Second-price Auctions', in M. Klush and F. Zambonelli (eds) *Lecture Notes in Artificial Intelligence, Vol. 2812*, Berlin: Springer, pp. 154-65.

Brandt, F. (2002), 'A Verifiable Bidder-resolved Auction Protocol', in *Proceedings of the Workshop on Deception, Fraud and Trust in Agent Societies, The International Conference on Autonomous Agents and Multiagents (AAMAS)*, Bologna: AAMAS, pp. 18-25.

Burt, R.S. (1982), *Toward a Structural Theory of Action. Network Models of Social Structure, Perception, and Action*, New York: Academic Press.

Buskens, V. (1998), 'The Social Structure of Trust', *Social Networks* 20(3): 265-98.

Carley, K. (1991), 'A Theory of Group Stability' *American Sociological Review* 56(3): 331-54.

Castelfranchi, C. and Falcone, R. (1998), 'Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification', in *Proceedings of the 3rd International Conference of Multi-Agent Systems (ICMAS'98)*, 3-7 July, Paris: ICMAS, pp. 72-79.

Castelfranchi, C. and Falcone, R. (2000a), 'Social Trust: A Cognitive Approach' in C. Castelfranchi and Yao-Hua Tan (eds) *Trust and Deception in Virtual Societies*, Dordrecht: Kluwer Academic Publishers, pp. 55-90.

Castelfranchi, C. and Falcone, R. (2000b) 'Trust Is Much More Than Subjective Probability: Mental Components and Sources of Trust, in *Proceedings of the 33rd Hawaii International Conference on System Sciences* Vol. 6, Maui, 4-7 January, IEEE Computer Society, <http://www.computer.org/proceedings/hicss/0493/04936/04936008abs.pdf>, accessed 17 Apr. 04.

Cerami, E. (2002) *Web Services Essentials*, Sebastopol CA: O'Reilly & Associates.

Cohen, P.R. and Levesque, H.J. (1990), 'Intention is Choice with Commitment', *Artificial Intelligence* 42(2-3): 213-61.

Dasgupta, P. (1988), 'Trust as a Commodity', in D. Gambetta (ed.) *Trust: Making and Breaking Cooperative Relations*, Oxford: Blackwell, pp. 49-72.

Deitel, M.H., Deitel, P.H. and Nieto, T.R. (2001), *E-business and E-commerce: How to Program*, Princeton NJ: Prentice Hall.

Dellarocas, C. (2002), 'Towards Incentive-compatible Reputation Management' in *Proceedings of the Workshop on Deception, Fraud and Trust in Agent Societies, The International Conference on Autonomous Agents and Multiagents (AAMAS)*, Bologna: AAMAS, pp. 26-40.

Durfee, E.H. (1999), 'Practically Coordinating' *AI Magazine* 20(1): 99-116.

eBay (2003), <http://www.ebay.com>, accessed 17 Apr. 04.

Falcone, R. and Castelfranchi, C. (2001), 'Social Trust: A Cognitive Approach' in C. Castelfranchi and Y.-H. Tan (eds) *Trust and Deception in Virtual Societies*, Dordrecht: Kluwer Academic Publishers, pp. 55-90.

Falcone, R., Singh, M.P. and Tan, Y. (eds) (2001), *Trust in Cyber-societies, Integrating the Human and Artificial Perspectives, Lecture Notes in Computer Science, Vol. 2246*, Berlin: Springer-Verlag.

Foster, I. and Kesselman, C. (eds) (1998), *The Grid, Blueprint for a New Computing Infrastructure*, San Mateo CA: Morgan Kaufmann Inc.

Foundation for Intelligent Physical Agents (2001), 'Abstract Architecture Specification, Version j', <http://www.fipa.org/specs/fipa00001/>, accessed 17 Apr. 04.

Gambetta, D. (1998), 'Can We Trust Trust?' in D. Gambetta (ed.) *Trust: Making and Breaking Cooperative Relations*, Oxford: Blackwell, pp. 213-37.

Gong, L. (2002), 'Tutorial: Peer to Peer Networks in Action', *IEEE Distributed Systems Online* 3(1),

Sarvapali Ramchurn and Nicholas Jennings

<http://dsonline.computer.org/0201/ic/w102gei.htm>, accessed 17 Apr. 04.

- Grandison, T. and Sloman, M. (2000), 'A Survey of Trust in Internet Applications', *IEEE Communications Surveys and Tutorials* 4(4): 2-16.
- He, M., Jennings, N.R. and Leung, H. (2003), 'On Agent-mediated Electronic Commerce', *IEEE Trans on Knowledge and Data Engineering*, 15(4): 985-1003.
- Herzberg, A., Mass, Y., Michaeli, J., Naor, D. and Ravid, Y. (2000), 'Access Control Meets Public Key Infrastructure, or: Assigning Roles to Strangers', in *Proceedings of the IEEE Symposium on Security and Privacy*, New York: IEEE Computer Society, pp. 2-4.
- Hsu, M. and Soo, V. (2002), 'A Secure Multi-agent Vickrey Auction Scheme', in *Proceedings of the Workshop on Deception, Fraud and Trust in Agent Societies, The International Conference on Autonomous Agents and Multiagents (AAMAS)*, Bologna: AAMAS, pp. 86-91.
- Jennings, N.R. (1993), 'Commitments and Conventions: The Foundation of Coordination in Multi-agent Systems', *The Knowledge Engineering Review* 8(3): 223-50.
- Jennings, N.R. (2001), 'An Agent-based Approach for Building Complex Software Systems', *Communications of the ACM* 44(4): 35-41.
- Jennings, N.R., Faratin, P., Lomuscio, A.R., Parsons, S., Sierra, C. and Wooldridge, M. (2001), 'Automated Negotiation: Prospects, Methods and Challenges', *International Journal of Group Decision and Negotiation* 10(2): 199-215.
- Jurca, R. and Faltings, B. (2003), 'An Incentive Compatible Reputation Mechanism', in *Proceedings of the IEEE Conference on E-Commerce CEC03*, Newport Beach CA, 24-27 June, New York: IEEE Computer Society, pp. 285-92.
- Kephart, J.O. and Chess, D.M. (2003), 'The Vision of Autonomic Computing', *IEEE Computer* 36(1): 41-50.
- Kollock, P. (1999), 'The Production of Trust in Online Markets', E. J. Lawler, M. Macy, S. T. Thyne, and H. A. Walker (eds) *Advances in Group Processes Vol. 16*, Greenwich CT: JAI Press, pp. 99-124.
- Kraus, S. (2001), *Strategic Negotiation in Multi-Agent Environments*, Cambridge MA: MIT Press.
- Mass, Y. and Shehory, O. (2001), 'Distributed Trust in Open Multi-agent Systems' in R. Falcone, M. Singh and Y.-H. Tan (eds) *Trust in Cyber-societies*, Berlin/Heidelberg: Springer-Verlag, pp. 159-73.
- Molm, L.D., Takahashi, N. and Peterson, G. (2000), 'Risk and Trust in Social Exchange: An Experimental Test of a Classical Proposition', *American Journal of Sociology* 105(5): 1396-1427.
- Mui, L. Mohtashemi, M. and Halberstadt, A. (2002), 'A Computational Model of Trust and Reputation for e-Business', in *35th Hawaii International Conference on System Science (HICSS 35 CDROM)*, IEEE Computer society (online publication),
<http://csdl.computer.org/comp/proceedings/hicss/2002/1435/07/14350188.pdf>
- Mukherjee, R., Banerjee, B., and Sen, S. (2001), 'Learning Mutual Trust', in R. Falcone, M.P. Singh and Y. Tan (eds) *Trust in Cyber-societies, Integrating the Human and Artificial Perspectives, Lecture Notes in Computer Science, Vol. 2246*, Berlin: Springer, pp. 145-58.
- Panzarasa, P., Jennings, N.R. and Norman, T. (2001), 'Social Mental Shaping: Modelling the Impact of Sociality on the Mental States of Autonomous Agents', *Computational Intelligence* 4(17): 738-82.
- Poslad, S., Calisti, M. and Charlton, P. (2002), 'Specifying Standard Security Mechanisms in Multi-agent Systems', in *Proceedings of the Workshop on Deception, Fraud and Trust in Agent Societies, The International Conference on Autonomous Agents and Multiagents (AAMAS)*, Bologna: AAMAS, pp. 122-27.
- Prietula, M. (2000), 'Advice, Trust, and Gossip among Artificial Agents', in A. Lomi and E. Larsen (eds) *Simulating Organizational Societies: Theories, Models and Ideas*, Cambridge MA: MIT Press, pp. 141-80.

Sarvapali Ramchurn and Nicholas Jennings

- Pynadath, D. and Tambe, M. (2002), 'Multi-agent Teamwork: Analyzing Key Teamwork Theories and Models', in C. Castelfranchi and L. Johnson (eds) *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems, Vol. 2*, New York: ACM Press, pp. 873-80.
- Resnick, P. and Zeckhauser, R. (2002), 'Trust among Strangers in Internet Transactions: Empirical Analysis of e-Bay's Reputation System', in M.R. Baye (ed.) *The Economics of the Internet and E-commerce, Advances in Applied Microeconomics, Vol. 11*, Amsterdam: Elsevier Science, pp. 127-57.
- Rosenschein, J. and Zlotkin, G. (1994), *Rules of Encounter: Designing Conventions for Automated Negotiation among Computers*, Cambridge MA: MIT Press.
- Russell, S. and Norvig, P. (1995), *Artificial Intelligence: A Modern Approach*, New York: Prentice Hall.
- Sabater, J. and Sierra, C. (2002), 'Regret: A Reputation Model for Gregarious Societies', in C. Castelfranchi and L. Johnson (eds) *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*, New York: ACM Press, pp. 475-82.
- Sadeh, N. (2002), *M-Commerce: Technologies, Services, and Business Models*, Chichester: Wiley Computer Publishing.
- Sandholm, T. (1999), 'Distributed Rational Decision Making', in G. Weiss (ed.) *Multi-Agent Systems: A Modern Approach To Distributed Artificial Intelligence*, Cambridge MA: MIT Press, pp. 201-58.
- Savage, L. (1954), *The Foundations of Statistics*, Hoboken NJ: John Wiley and Sons.
- Schillo, M., Funk, P. and Rovatsos, M. (2000), 'Using Trust for Detecting Deceptive Agents in Artificial Societies', *Applied Artificial Intelligence, Special Issue on Trust, Deception, and Fraud in Agent Societies* 14(8): 825-48.
- Schmeck, H., Ungerer, T. and Wolf, L.C. (eds) (2002), *Trends in Network and Pervasive Computing ARCS 2002, International Conference on Architecture of Computing Systems, Karlsruhe, Germany, April 8-12, Proceedings* (Lecture Notes in Computer Science, Vol. 2299), Berlin: Springer-Verlag.
- Sen, S. (1996), 'Reciprocity: A Foundational Principle for Promoting Cooperative Behavior among Self-interested Agents', in *Proceedings of the Second International Conference on Multi-agent Systems*, Menlo Park CA: AAAI Press, pp. 322-29.
- Sen, S., Biswas, A. and Debnath, S. (2000), 'Believing Others: Pros and Cons', in *Proceedings of the Fourth International Conference on Multi-Agent Systems*, Boston, 10-12 July, pp. 279-86.
- Sen, S. and Dutta, P.S. (2002), 'The Evolution and Stability of Cooperative Traits', in C. Castelfranchi and L. Johnson (eds) *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems, Vol. 3*, New York: ACM Press, pp. 1114-20.
- Sen, S. and Sajja, N. (2002), 'Robustness of Reputation-based Trust: Boolean Case', in C. Castelfranchi and L. Johnson (eds) *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems, Vol. 1*, New York: ACM Press, pp. 288-93.
- Simon, H.A. (1996), *The Sciences of the Artificial*, Cambridge MA: MIT Press, (3rd edition).
- Singh, M.P., Yu, B. and Venkatraman, M. (2001), 'Community-based Service Location', *Communications of the ACM* 44(4): 49-54.
- von Neuman, J. and Morgenstern, O. (1944), *The Theory of Games and Economic Behaviour*, Princeton NJ: Princeton University Press.
- Vulkan, N. (1999), 'Economic Implications of Agent Technology and e-Commerce', *The Economic Journal* 109(453): 67-90.
- Witkowski, M., Artikis, A. and Pitt, J. (2001), 'Experiments in Building Experiential Trust in a Society of Objective-trust Based Agents', in R. Falcone, M.P. Singh and Y. Tan (eds) *Trust in Cyber-societies, Integrating the Human and Artificial Perspectives, Lecture Notes in Computer Science, Vol. 2246*, Berlin: Springer, pp. 111-32.

Sarvapali Ramchurn and Nicholas Jennings

Wooldridge, M. (2002), *An Introduction to Multi-Agent Systems*, Chichester: John Wiley & Sons.

Wu, D.J. and Sun, Y. (2001), 'The Emergence of Trust in Multi-agent Bidding: A Computational Approach', in *Proceedings of the 34th Hawaii International Conference on System Sciences (HICSS-34, CD ROM)*, Vol. 1, IEEE Computer Society Press.

Yager, R.R., Kacprzyk, J. and Fedrizzi, M. (1994), *Advances in the Dempster-Shafter Theory of Evidence*, Chichester: John Wiley.

Yamagishi, T., Cook, K. and Watabe, M. (1998), 'Uncertainty, Trust, and Commitment Formation in the United States and Japan', *American Journal of Sociology*, 104(1): 165-94.

Yu, B. and Singh, M.P. (2002a), 'Distributed Reputation Management for Electronic Commerce', *Computational Intelligence*, 18(4): 535-49.

Yu, B. and Singh, M.P. (2002b), 'An Evidential Model of Reputation Management', in C. Castelfranchi and L. Johnson (eds) *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems, Vol. 1*, New York: ACM Press, pp. 294-301.

Yu, B. and Singh, M.P. (2003), 'Social Networks and Trust: Searching Social Networks', in *Proceedings of the 2nd International Joint Conference on Autonomous and Multi-Agent Systems*, New York: ACM Press, pp. 65-72.

Zacharia, G. and Maes, P. (2000), 'Trust through Reputation Mechanisms', *Applied Artificial Intelligence*, 14(9): 881-908.

