Applied Artificial Intelligence, 18:833–852, 2004

ISSN: 0883-9514 print/1087-6545 online
DOI: 10.1080/08839510490509045

Taylor & Francis
Taylor & Francis Group

DEVISING A TRUST MODEL FOR MULTI-AGENT INTERACTIONS USING CONFIDENCE AND REPUTATION

SARVAPALI D. RAMCHURN and
NICHOLAS R. JENNINGS
School of Electronics and Computer Science,
University of Southampton, Southampton,
United Kingdom

CARLES SIERRA and
LLUIS GODO
Artificial Intelligence Research Institute,
Bellaterra, Spain

In open environments in which autonomous agents can break contracts, computational models of trust have an important role to play in determining who to interact with and how interactions unfold. To this end, we develop such a trust model, based on confidence and reputation, and show how it can be concretely applied, using fuzzy sets, to guide agents in evaluating past interactions and in establishing new contracts with one another.

Agents generally interact by engaging in some form of negotiation process which results in them making commitments to (contracts with) one another to carry out particular tasks (Jennings et al. 2001). However, in most realistic environments, there is no guarantee that a contracted agent will actually enact its commitments (because it may defect to gain higher utility or because there is uncertainty about whether the task can actually be achieved). In such situations, computational models of trust (here defined as the positive expectation that an interaction partner will act benignly and cooperatively in situations where defecting would prove more profitable to itself [Dasgupta 1998]) have an important role to play. First, to help determine the most reliable interaction partner (i.e., those in which the agent has the highest trust). Second, to influence the interaction process itself (e.g., an agent's negotiation stance may vary according to the opponent's trust level). Third, to define the

Address correspondence to Sarvapali D. Ramchurn, School of Electronics and Computer Science, University of Southampton, SO17 1BJ, City, United Kingdom. E-mail: sdr01r@ecs.soton.ac.uk

set of issues that need to be settled in the contract (i.e., the higher the trust, the more that can be left implicit in the contract).

Generally speaking, agent interactions go through three main phases: (i) a negotiation dialogue during which the terms of the contract are agreed upon and agents assign an expected utility to the contract, (ii) an execution phase during which there are opportunities for the contracted agent to defect, and (iii) an outcome evaluation phase where the client agent assesses the outcome of the task and finally derives some utility. In the cases where an agent has an incentive to defect, the client agent can judge whether the contractor is trustworthy by assessing its performance, relative to the initially contracted agreement, given its *perception of the contract and the context*. Thus, the trust value for a specific agent for a specific task needs to take into account the potential utility loss or risk[1] (associated with the task in question) in a contract given information about the context in which the contract is enacted (Marsh 1994). This follows from the fact that cooperating under high potential losses shows greater trustworthiness than otherwise (Yamagishi et al. 1998).

Trust values, thus devised, can guide future contract negotiations in order to ensure that guarantees are provided against losses. Thus, if trust is sufficiently high, the contracted agent is deemed reliable. This means less time can be spent looking for potential contractors, negotiating about the minute guarantees present in the contract and, accordingly, giving more freedom to the contracted agent to enact its part of the deal. Conversely, when trust is low, the agents may spend a significant time specifying the guarantees associated with a contract or, if possible, avoiding future interactions with such agents.

Given this background, a number of computational models of trust have been developed (see [Ramchurn et al. 2004] for more details). In Marsh (1994) for example, trust is taken to be a value between -1 and 1 that is calculated by taking into account risk in the interaction and the competence level of an interaction partner. However, the calculation of risk is not given and the model does not take into account past experience and reputation values of the contracted agent. In Sabater and Sierra (2002), reputation symbolizes trust and competence levels are gathered from the social network in which the agents are embedded. The main value of this model lies in showing how reputation can be used to guide an agent's negotiation stance, but the evaluation of direct interactions is overly simple (disregarding the context).

In general, extant trust models fail to capture the individuality of an agent in assessing the reliability of an interaction partner. Most models also neglect the fact that agents interact according to the norms and conventions determined by the society or environment within which they are situated (Esteva et al. 2001). To this end, this paper develops a computational model of trust that rectifies these shortcomings.

By taking into account its past experience (from direct interactions) and information gathered from other agents (indirect interactions), an agent can

build up beliefs about how trustworthy a contracted agent is likely to be in meeting the expected outcomes of particular contract issues (e.g., delivering goods on time or delivering high quality goods). In this respect, we conceive of two ways of assessing trustworthiness: (i) *confidence* derived (mainly) from analyzing the result of previous interactions with that agent and (ii) *reputation* acquired from the experiences of other agents in the community through gossip or by analyzing signals sent by an agent. Both measure the same property; that is, the agent's believed reliability in doing what it says it will regarding particular issues of a contract. Here these measures are modeled using fuzzy sets to give agents a robust means of assessing the extent to which their interaction partners satisfy the issues of a contract. In particular, we advance the state of the art in the following ways. First, we delineate and computationally model context information, confidence measures, and risk in agent interactions. Second, we show how a measure of trust can be derived from the latter concepts and reputation information. Finally, we show how the trust measure can guide the choice of interaction parties, the stance that is taken during negotiation, and the issues that need to be agreed upon in a contract.

THE TRUST MODEL

In this section, we describe the trust model.[2] We first give the basic definitions that we will use in the rest of the paper. Using these definitions, we model confidence, reputation, and norms. We then show how to combine these measures to compute appropriate trust values.

Basic Notions

Let Ag be the society of agents noted as $\alpha, \beta, \ldots \in Ag$. A particular group of agents is noted as $G \subseteq Ag$ and each agent can only belong to one group.[3] Groups can be ordered in terms of the power relationships that hold between them. This captures the notion of authority that one group may hold over another one. The power relationship is modeled as a total pre-order on the groups. Thus, if G_i has at least as much power as G_j, this is noted as $G_i \succeq_p G_j$. We conceive that agents within each group have a set of similar norms (e.g., all retailers in the U.K. agree to a 14-day return policy on all items they sell or all retailers in Spain close on Sunday). τ denotes a totally ordered set of time points (sufficiently large to account for all agent interactions) noted as $t_0, t_1, \ldots$, such that $t_i > t_j$ if and only if $i > j$.

Contracts

Contracts are agreements about issues and the values these issues should have. Let $X = \{x_1, x_2, \ldots, x_n\}$ be the set of potential issues to include in a contract, and the domain of values taken by an issue x be noted as D_x

(for simplicity we assume that all D_x are an interval of real numbers $\mathbb{R}$). We will note that issue x takes the value $\upsilon \in D_x$ as $x = \upsilon$. Thus, a particular contract, O, is an arbitrary set of issue-value assignments noted as $\mathcal{O} = \{x_1 = \upsilon_1,\ x_2 = \upsilon_2, \ldots, x_n = \upsilon_n\}$, where $x_i \in X, \upsilon_i \in D_{x_i}$ We denote by O the set of potential contracts. We will also note the set of issues involved in contract O as $X(O) \subseteq X$. Given an agreed contract, two or more agents all have a subset of the contract to enact. Each subset of the contract allocated to an agent is superscripted by the respective agent identifier such that, for example, in a contract O between α and β, $O^\alpha \cup O^\beta = O$. An agent, α, has a utility function for contracts, noted as $U^\alpha : \mathcal{O} \rightarrow [0, 1]$, and for each issue $x \in X(O)$ in a contract noted as $U^\alpha_x : D_x \rightarrow [0, 1]$. In this work, we will define the utility of a contract, for an agent, as a weighted aggregation of the utilities of the individual issues as shown next (note this assumes that issues are independent): $U^\alpha(O) = \Sigma_{x \in X(O)^{\omega_x}} . U^\alpha_x(\upsilon_x)$, where $\Sigma\omega_x = 1$ and $\upsilon_x \in D_x$ is the value taken by the issue $x \in X(O)$. We also define a similarity function for an issue x as $Sim_x : D_x \times D_x \rightarrow [0, 1]$, which determines how similar two values for the same issue are. Such a function is required to have the following two properties (i) $Sim_x(\upsilon, \upsilon) = 1$ (i.e., reflexivity) and (ii) $Sim_x(\upsilon_1,\ \upsilon_2) = Sim_x(\upsilon_2, \upsilon_1)$ (i.e, symmetry). A global similarity function over contracts Sim can also be applied over two contracts having the same issues by aggregating the similarity values of each issue in each contract such that: $Sim(O, O') = \Sigma_{x \in X(O) = X(O')} \omega_x \times Sim_x (\upsilon_x, \upsilon'_x)$, where w_x is the weight of issue x and $\Sigma w_x = 1$. The weight w_x can be the same as those used in the utility function so that contracts deemed similar also have similar utilities for the agent concerned.

We consider that agents, whether from the same group or from different groups, invariably interact within some electronic institution (Esteva et al. 2001) which specifies and (or) restricts (some) issue-value assignments of contracts through a set of norms (see the next sub-section). Naturally, each institution may also specify different rules.

Rules Dictating Expected Issue-Value Assignments

The agreed contract provides a clear statement of what is expected with respect to each issue. However, the social setting in which the interaction takes place may also give rise to expectations which are not explicitly stated in the contract itself. For example, a buyer agent α from country A might expect seller agent β from country B to deliver goods nicely wrapped in gift paper as opposed to in a carton box. This clause may not have been specified in the contract as it is a norm in the client's group that goods must be nicely wrapped. Thus, at execution time, an agent may fail to satisfy another's (contracted or not) expectations because (i) it is not able to meet the expectations, (ii) it is not willing to meet the expectations, or (iii) it is not aware of the unspecified expectations. In any case, the satisfaction or not of these expectations *directly* impacts on the trust the agent has in its opponent (Molm et al.

2000). If a satisfactory reason is given for poor performance, the trust value may not be modified, but this is not considered here.

Against this background, we take into account the three basic sets of norms[4] that can be sources of unspecified expectations[5]: (i) *social rules*, noted as *SocRules*, that all agents in the society Ag possess in common, (ii) *group rules*, noted as *GroupRules* (G), that all agents within a particular group $G \subseteq Ag$ have in common, and (iii) *institutional rules*, noted as *InstRules*, that agents α and β interacting within a particular electronic institution must abide by. In the case of group rules, there is no guarantee that agents from different groups, having different norms, will satisfy their interaction partner's group rules. On the other hand, the conclusions of institutional rules are guaranteed by the institution (e.g., price p has to be paid, seller has to give goods). This guarantee is normally specified through a penalty which must be paid (by the rule breaker) if the rule is not respected. In more detail, rules of all types allow an agent to infer expected issue-value assignments from a contract. Here the rules will be written in the following way: **IF** $x_1 = v_1$ *and* $x_2 = v_2$ *and* ... *and* $x_m = v_m$ **Then** $x = v$, meaning that if $(x_i = v_i) \in O$ for all $i = 1, \ldots, m$, then issue x's value is expected to be equal to v. An example of such a rule would be: **IF** price = £100 and qos = 8 **Then** anti-DoS = 10, which means that the if the price of the an telecommunication line (bought from some Internet Service Provider [ISP]) is a hundred pounds, and the quality of service guarantee (qos) of the ISP is eight (i.e., high in this context), then it is expected that the ISP will provide an anti denial-of-service (DoS) on the line. We assume that x does not appear in the premise of the rule (otherwise this could lead to some rules being unsatisfiable). We note by *Rules* the set of all possible rules written using the above syntax[6] over the set X of issues and corresponding domains of values. The rules an agent abides by will depend on the group it belongs to and the other rules implied by the institution within which it is interacting with others.

Given a contract O proposed by α to β, where $\alpha \in G_1$ and $\beta \in G_2$, we can now devise the set of all of α's (or β's) expectations (unspecified and specified) about the values of the issues in the contract. The unspecified expectations due to the social setting, O^{α}_{exp}, of issue-value assignments from O is the set of all conclusions of the rules of agent α, *Rules* (α) = *SocRules* $\cup$ *Group Rules* (G_1) and *InstRules* (that apply to α and β), that have their premise satisfied by the equalities in the contract O. The complete expanded contract from α's point of view is therefore defined as $O^{\alpha}_{+} = O \cup O^{\alpha}_{exp}$ (the latter will be different from β's expanded contract, O^{β}_{+}, if β's group has different rules that apply to the issues of O).

The issues contained in the expanded contract may vary (for the same contract O), depending on the group and institutional rules that apply at the time the agents make an agreement. This is because an agent may interact under different institutions (having different institutional norms) or an agent

may decide (be allowed) to switch groups to one that has different norms (power) from its original group. Given the expanded contract, an agent may then decide to trust its opponent depending on its prior knowledge of its opponent's performance. In the following text we model this in more detail.

Interaction History and Context

In order to try and predict the future performance of an agent, it is important to analyze its interaction history in terms of both the outcomes of interactions and the norms that prevailed in each past interaction. In more detail, the interaction history of an agent α, intending to interact with an agent β, can be viewed as consisting of a list of elements with four main components: (i) α's agreed contract O with β and the outcome of the enactment of the contract O' by β and α (i.e., a list of pairs of (O, O') form the contracting history), (ii) $Rules(\alpha)$ that α had to abide by for the contract (at time t_i when the contract was signed), (iii) *InstRules* that both α and β had to abide by in a given institution, and (iv) α's utility function (at time t_i) for the contract issues for which it hired β. Each element in an interaction history is therefore represented as: $c_i = \langle \alpha, \beta, O, O', \{U_x^\alpha\}_{x \in X(O)}, Rules(\alpha), InstRules, t_i \rangle$, and the interaction history as $CB = \{c_1, c_2 \ldots\}$. We will note by $CB_{\alpha,\beta} \subseteq CB$, the history containing all interactions between α and β.

For each new interaction between α and β, α will need to consider the interaction history as well as the currently prevailing rules and its current utility function in order to predict the behavior of β (as will be shown later). Thus we define as α's *current* context[7] within which a new contract is negotiated with an agent β and executed as the set: $\Sigma_{\alpha,\beta} = \langle CB_{\alpha,\beta}, \{U_x^\alpha\}_{x \in X}, Rules(\alpha), t_c \rangle$, where t_c represents the current time. We assume that the agents will have agreed between them (through negotiation or by one partner imposing the institutional rules) which institution will guide their interactions and this will imply a given set of rules *InstRules* applying over the interaction.[8]

Every time a new contract is agreed and enacted, it is added as a new element to CB in order to update the context of the agent. Moreover, all the rules, including the *InstRules*, will be recorded in the interaction history after the interaction is completed. Thus, this context can be dynamic for a number of reasons (apart from the history being updated with new elements). First, an agent may change groups such that its group rules might change and, consequently, so will its expectations. Second, an agent may interact with the same partners within different institutions (e.g., buying from a seller in England and buying from the same seller in Spain where different trade rules or laws apply). Third, the interacting agents might change their utility functions over time such that they value an issue differently at different points in time (e.g., a travel package may be worth more in summer than in winter).[9]

In the following sections, we use information derived from the context in order to define and evaluate the agent's trust in its opponent's enactment of

the contractual terms. We will differentiate between the trust derived from personal knowledge about an agent (confidence) and that derived from information about the agent gathered from other agents in the society (reputation). In the next section, we focus on defining confidence (i.e., the personal aspect of trust) and later combine it with reputation (which is based on the confidence of other agents) to get an overall notion of trust.

Confidence

We will define confidence as follows: α's confidence is an issue x handled by β is a measure of certainty (leading to trust), based on evidence from past direct interactions with β, which allows α to expect a given set of utility deviation values to be achieved by β for x.[10]

Thus if α has a high degree of confidence with respect to x being well enacted or not by β, then the interval of utility deviation values expected by α and β will be relatively *small* (conversely the set is large if confidence is low). This set of utility deviation values may bring more utility than expected (i.e., a high confidence in β being "good") or less utility than expected (i.e., a high confidence in β being "bad"). We initially consider confidence on a per-issue basis given that agents may be more reliable in satisfying some issues than others. These measures of imprecision on an opponent's behavior are not strictly probabilistic in nature since they may involve a subjective appreciation of performance as well (e.g., how "bad" or "good" the delivery time of goods is for a buyer might not be precisely defined and this perception might also vary over time depending on the agent's preferences). Given this, we choose a fuzzy set-based approach to relate confidence levels with expected utility deviations for issues.

Confidence Levels

In this work, the behavior of an agent regarding the fulfillment of a term (i.e., an issue-value pair) in a contract is perceived in terms of the variations on utility between the signed value for the issue and the enacted one. These utility variations are then sensed over multiple interactions to build a picture of the agent's performance over time. Here, we take the stance that fuzzy sets have their domains specified over "absolute" variations on utility ΔU, rather than on relative variations (i.e., relative to the utility of the value signed for the issue).[11] Thus, we consider that $\Delta U \in [-1, 1]$ (recall that utility values belong to the interval [0,1]).

Specifically, we assume that agents share a (small) set $\mathcal{L} = \{L_1, L_2, \ldots, L_k\}$ of linguistic labels to qualify the performance of an agent on each issue. In what follows, we will use the basic set $\mathcal{L} = \{\text{Bad, Average, Good}\}$. We believe these labels provide an adequate means for an agent to express its view on the *possible* (approximate) utility deviations, gains, or losses, in the

TABLE 1 Possible Different Meanings of the Labels for Three Agents When Applied to the Issue "Delivery"

Label/Agent	α	β	γ
Bad	Late	Very late	Too late
Average	On time	Just in time	Right time
Good	Early	Very early	Early enough

executed contract with respect to the utility of the contractually signed values. For example, each agent could understand the labels "Bad," "Average," and "Good" for the issue "delivery" in different ways according to their ontology (as shown in Table 1).[12] Thus, using Table 1, agent α can translate a "Very Late" rating from an agent β as *Late* (since they both equate to "Bad") and "Right time" from γ as "On time" (since they both equate to "Average"). In more detail, we model the meaning of a label L by a fuzzy set on the domain of utility deviations $\Delta U = [-1, 1]$, specified by its membership function $\mu_L(u) : [-1, 1] \rightarrow [0, 1]$. Examples of membership functions[13] for the above set of labels are given in Figure 1.

Thus, agent α's confidence level is defined as the membership level, measured over [0,1], of the behavior of a particular agent β with respect to an issue x to a linguistic term L, noted as $C(\alpha, \beta, x, L)$. In the remainder of the paper, we will avoid the agent identifier(s) wherever this is unambiguously defined by the context. Therefore, the cut of the fuzzy set defined by $C(x, L)$ represents a range (on the horizontal axis) of values, $E\Delta U_c(x, L) = \{\delta u \in [-1, 1] \mid \mu_L(\delta u) \geq C(x, L)\}$, that is understood as the range of expected utility deviations at execution time on issue x by agent β. For instance, α may express its belief that β is "Good" to a confidence level 0.6 in fulfilling the contractual values on price, "Average" to a level of 0.25, and "Bad" to a level of 0. This would mean that α expects the utility devi-

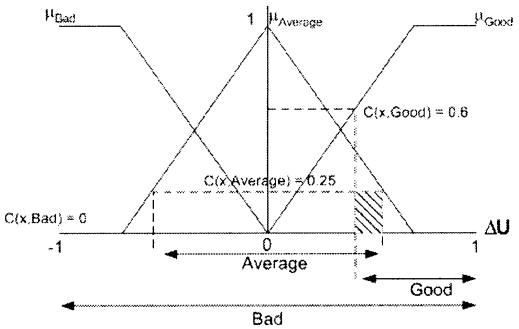

FIGURE 1 Shapes of membership functions in different labels and ranges supporting confidence levels in "Good" (0.6), "Average" (0.25) and "Bad" (0). The shaded region indicates the range over which the sets "Good" and "Average" intersect. The base of this shaded region is the set of expected values of ΔU.

ation to lie within the range of values which support the confidence level of 0.6 for "Good," 0.25 for "Average," and 0 for "Bad." This is shown on Figure 1.

Evaluating Confidence

In order to obtain confidence levels for different labels, we first need to calculate the range of utility variations expected in the issue. This expected range can be obtained by considering the utility changes that have been observed in past interactions.[14] While the size of samples of ΔU_x will naturally determine the accuracy of the model, the temporal range of samples taken (i.e., a window over the latest interactions) will determine how up-to-date the model is in determining the current nature of the opponent. Therefore, we propose two ways of eliciting the confidence level from a probability distribution that minimize computational complexity. The first considers using the confidence interval of a normal distribution,[15] while the second uses a time dependent mean.

Therefore, given a context $\Sigma_{\alpha,\beta}$ and a proposed (not yet agreed) contract O, for each issue x in $X(O)$, we can estimate, from the history of past interactions, a probability distribution P of α's utility variation $\Delta U_x \in [-1, 1]$ (negative or positive) relative to issue x (we will avoid the agent identifier in the utility function since this is clear from the context). Values of ΔU_x correspond to the possible differences between the utility $U_x^{\alpha}(v)$ of the agreed value $(x = v) \in O$ and the utility $U_x^{\alpha}(v')$ of the (unknown) final value $(x = v')$ in the executed contract O' (i.e., $\Delta U_x = U_x^{\alpha}(v') - U_x^{\alpha}(v)$). Then we can say that the agent α has a certain *risk* with issue x when it estimates that $1 \geq q > 0$ where q is the probability that $\Delta U_x < 0$. Of course, the more negative the mean, $\overline{\Delta U_x}$, of this probability distribution (i.e., the higher the expected utility loss), the higher the risk, and the more positive this mean is, the lower the risk (i.e., the lower the expected utility loss).

Thus, to calculate the confidence levels in each of the issues concerned, we first need to estimate the probability distribution of P. This has to be done both for those issues x appearing in O and those in the expanded contract $O_+ = O \cup O_{exp}$, resulting from the application of the rules in the current context. We have to do so analogously with the contracts in the precedent cases of the interaction history CB of the current context. However, if we assume that the proposed contract is signed such that the norms of the institution *InstRules* under which the agents (α and β) are operating are fully enforced (i.e., penalties, matching the utility loss on an issue, have to be paid by the agent which does not respect the norms which regiment the performance on the issue), then the risk is zero[16] for those (groups of) issue-value assignments insured by institutional norms. This is the case even though the inference from previous interactions may suggest that the agent would defect. In such cases, we remove all these insured issues from the analysis. In the same way, if in an element of the interaction history, an

issue's enactment was guaranteed by the institution under which the agents interacted at the time, we remove it from the sample of elements being analyzed for that issue. This procedure avoids us incrementing trust when an institution has guaranteed good behavior in the past (since risk is zero in such cases).

In cases where the interacting agents have made contracts with very different values (e.g., trading a car involves more utility transfer than trading a phone), it might be important to assess the history given the similarity of past cases. This prevents an agent from building trust using low valued transactions in order to be trusted on high value transactions which it can defect from in order to reap a large profit (e.g., on eBay [Resnick and Zeckhauser 2002]). To this end, the similarity functions introduced earlier can be used to filter the cases which are most relevant to the current contract O as follows:

$$CBS_{\alpha,\beta}(O,x) = \left\{ c \in CB_{\alpha,\beta} \middle| \begin{array}{c} c = \langle \alpha, \beta, O_1, O_2, \{U_y^{\alpha}\}_{y \in X(O)}, Rules(\alpha), \\ InstRules, t \rangle, \\ (x = v) \in O, (x = v') \in O_1, \\ Sim(O, O_1) > \lambda_O, Sim_x(v, v') > \lambda_i \end{array} \right\}$$

where $CBS_{\alpha,\beta}(O, x)$ is the set of cases to consider in the assessment of β's behavior with respect to x on a given contract. The rule above selects those cases for which the similarity level of the contracts, as well as the similarity of values for that issue, are above certain thresholds $\lambda_O \in [0,1]$ and $\lambda_x \in [0,1]$.

Now, assume we have defined the probability distribution P (obtained from $CBS_{\alpha,\beta}(O, x)$). From this distribution, different techniques (detailed later in this section) can be used to elicit a significantly representative interval $[\delta_1, \delta_2]$ for ΔU from which the confidence level can be obtained. To calculate confidence levels $C(x, L)$ for each label $L \in \mathcal{L}$, we want the interval $[\delta_1, \delta_2]$ to coincide as much as possible with the set of expected values $E\Delta U_c(x)$. Since this range is defined by the confidence levels of its limits, the procedure amounts to selecting the minimum confidence levels of the two limits for that label as: $C(x, L) = \min(\mu_L(\delta_1), \mu_L(\delta_2))$. We propose two ways to obtain the interval $[\delta_1, \delta_2]$:

1. Find the 95% (or higher) confidence interval of the mean of the distribution such that the probability that $(\delta_1 \leq \overline{\Delta U}_x \leq \delta_2)$ is equal to 0.95. This involves first approximating the distribution to a normal distribution by calculating the estimated sample variance $\hat{\sigma}^2$ of the distribution as well as the mean. Then the confidence interval can be obtained from the following equation: $\overline{\Delta U}_x \pm \frac{\hat{\sigma} \cdot l_{conf}}{\sqrt{N}}$, where $l_{conf} = 1.96$ for a 95% confidence and N is the sample size.

2. Instead of extracting a range of values, we choose an arbitrary value for δ_2 such that $\mu_L(\delta_2) = 1$ and then obtain δ_1 as:

$$\delta_1 = \sum_{c_i \in CB_{\alpha,\beta}} w_i \times (U_x^{\alpha}(\upsilon) - U_x^{\alpha}(\upsilon_i))$$

where $c_i = \langle \alpha, \beta, O, O', \{U_x^{\alpha}\}_{x \in X(O)}, Rules(\alpha), InstRules, t_i \rangle$ υ is the value taken by $x \in X(O)$ and υ_i is the value taken by $x \in X(O')$ and $w_i = \frac{Sim_x(\upsilon,\upsilon_i) \times \rho(t_c,t_i)}{\Sigma_{c_j \in CB_{\alpha,\beta}} Sim_x(\upsilon,\upsilon_j) \times \rho(t_c,t_j)}$. In the latter expression, t_c is the current time and $\rho: Time \times Time \rightarrow [0,1]$ is a time-weighing function which weighs the most recent contracts with a larger value. An instance of the time-weighting function is $\rho = \sin(\frac{\pi}{2} \cdot \frac{t_i}{t_c})$ where t_i is the time at which case i was recorded and t_c is the current time. Thus, δ_1 represents an expected value for the enacted contract given most cases with *similar* values for issue x (to the current value) as well as the most *recent* cases are given more importance in the evaluation of the mean.

We will assume that all agents in the society are able to evaluate their confidence in issues handled by their opponents and may transmit these measures to others. The transmission of such confidence then gives rise to the concept of *reputation*, which is described next and later combined with personal confidence measures.

Reputation

An agent's reputation is the perception of a group or groups of agents in the society about its abilities and attributes (Dasgupta 1998; Sabater and Sierra 2002). Several models of reputation have been developed to show how an agent can build up its trust in another by retrieving and aggregating information about the latter from other agents (e.g., [Yu and Singh 2002; Sabater and Sierra 2002; Zacharia and Maes 2000]). Thus, here we do not consider how this reputation information is gathered from the other agents in the society as there already exists several techniques to do this efficiently. Rather, we assume this information is simply available from a social network that structures the knowledge that each agent has of its neighbors and keeps track of past interactions (Sabater and Sierra 2002). This allows us to focus on representing reputation and combining it with confidence (as shown in the next sub-section) In this work, we specialize the definition of reputation to the following: α's estimate of β's reputation in handling an issue x is α's measure to certainty (leading to trust), based on the aggregation of confidence measures (for x) provided to it by other agents which have previously interacted with β which allows α to expect a given set of values to be achieved by β for x.

Hence, we assume that an agent α possesses a function *Rep:* $Ag \times X \times \mathcal{L} \rightarrow [0,1]$ where $Rep(\beta, x, L)$ represents α's view of agent β's reputation in handling issue x with respect to the qualifying label L (the name of the agent(s) will be omitted when the context unambiguously determines it). We also assume that the labels $L \in \mathcal{L}$ have their domain specified over the same range of utility deviations (i.e., $\Delta U \in [-1, 1]$ as explained earlier.

The reputation function generally aggregates confidence levels obtained from other agents in the environment (Sabater and Sierra 2002).[17] Here we connect to the REGRET reputation model by using the *witness reputation*.[18] The latter defines the reputation of an agent β as a sum of the confidence levels from other agents in the environment who have interacted in the past with β. Moreover, given our modeling of the society of agents in terms of groups and the power relationship holding between them, we use this power relationship to weigh the confidence levels sent (rather than using trust values as in REGRET). Thus, if group G_1 has more power than group G_2, then more weight is given to confidence levels transmitted from agents in group G_1 than to those from agents in G_2. This is because it is assumed that those agents that come from high power groups (e.g., legal institutions, government) have more credibility than those in low power groups (e.g., small companies, individuals). The witness reputation is therefore calculated by taking the minimum rating sent by agents in each group as: $Rep(\beta, x, L) = \Sigma_{G_i \in \mathcal{G}, \alpha \notin G_i} w_i \times \min_{\gamma \in G_i} C(\gamma, \beta, x, L)$ where $w_i > w_j$ iff $G_i \succeq_p G_j$.

As can be seen, reputation measures can be particularly useful to an agent that enters a system for the first time. This is because the agent would not have interacted with any other agents in its environment in the past. Therefore, it would not be possible for it to compute its confidence in them. Thus it can only use information that is supplied to it by other agents in the environment. However, such information may be liable to noise or may not be true if agents are lying. Also if no one else has interacted with an agent's opponent, then the agent can only take a guess at its opponent's reliability. In such circumstances, the agent can only learn from its direct interactions with other agents and compute its confidence measures from these interactions.

Given the above, using just confidence or just reputation values to compute the set of expected values for a given issue is often only useful in extreme situations. In the next section, we devise a measure that caters for all situations between these extremes and then after that we derive a trust measure from this.

Combined Confidence and Reputation Measures

Generally speaking, we consider that both confidence and reputation should be taken into account in order to come up with a set of expected values for an issue. We rely on a combination of both measures in order to balance both the societal view on an opponent and the personal view of the agent until the latter can be sure

that its own view is more accurate. To come to this conclusion, each agent will have its own threshold on the number of interactions needed to have this accurate measure. Therefore, given agent α's context $\Sigma_{\alpha,\beta}, \langle CB_{\alpha,\beta}, \{U_x^\alpha\}_{x\in X}, Rules(\alpha), t_c\rangle$, here we propose to define the threshold κ as $\kappa = \max(1, |CB_{\alpha,\beta}|/\theta_{min})$, where $|CB_{\alpha,\beta}|$, is the number of interactions of α with β and θ_{min} is the minimum number of interactions (successful negotiations and completed executions[19]) above which only the direct interaction is taken into account (Sabater and Sierra 2002).

Thus, we capture the combination of confidence and reputation measures through the function $CR : Ag \times X \times \mathcal{L} \rightarrow [0, 1]$, which is, in the simplest case, a weighted average of both kinds of degrees (as in the previous cases we omit references to the agent whenever possible):

$$CR(\beta, x, L) = \kappa \cdot C(\beta, x, L) + (1 - \kappa) \cdot Rep(\beta, x, L) \tag{1}$$

Given CR levels it is then possible to compute the expected values for an issue x and label L as: $E\Delta U_{cr}(\beta, x, L) = \{u \mid \mu_L^x(u) \geq CR(\beta, x, L)\}$, and then the intersection of the expected ranges for all the labels $L \in \mathcal{L}$ is: $E\Delta U_{cr}(\beta, x) = \cap_{L\in\mathcal{L}} E\Delta U_{cr}(x, L)$.

The assignment of CR values for all labels may not always be consistent (i.e., $E\Delta U_{cr}(\beta, x, L) = \emptyset$). This could happen, for example, if the agents in the environment do not hold the same view on one of their members (such that $Rep(\beta, x, L)$ does not define a consistent range). Our solution to this problem is the following: Whenever the intersection results in an empty set, we will iteratively not consider the label with the lowest confidence level, until a non-null range of values is obtained. This procedure equates to removing those decision variables that have the lowest importance in the set under consideration. Our solution ensures that a consistent intersection can be found in all possible cases.

As can be seen, the above range is defined in terms of the utility deviations rather than in terms of the values that the issue could take. However, at negotiation time, for example (as will be seen later), we might need to compute the expected values an issue could take, after execution of the contract, given an offered value v_0 for the issue. This requires transfering the expected utility deviations to the domain of the issue considered.[20] This can be computed in the following way:

$$EV_{cr}(\beta, x, v_0) = \{v \in D_x | U_x^\alpha(v) - U_x^\alpha(v_0) \in E\Delta U_{cr}(\beta, x)\} \tag{2}$$

Trust Measures

In our trust model, we use the combined degrees $\{CR(\beta, x, L)\}_{L\in\mathcal{L}}$, as given by Eq. 1, to define the interval of expected values $E\Delta U_{cr}(\beta,x)$ that

provides us with a maximum expected loss in utility $\Delta^{cr}_{loss}(x) = \sup(E\Delta U_{cr}(\beta, x))$. This maximum expected utility loss represents the risk that is involved in the interaction given knowledge acquired both from direct interactions and reputation and also from the norms of the environment. While the risk describes how much we expect to lose from an interaction, trust is the opposite of this given our initial definition. Thus we define trust as: $T(\beta, x) = \min(1, 1 - \Delta^{cr}_{loss}(x))$, where where T serves to describe trust in β for issue x based on both confidence in β and its reputation with respect to issue x.

Here, we choose to bound trust values[21] in the range [0,1], where 0 represents a completely untrustworthy agent (and corresponds to the maximum possible utility loss) and 1 represents a completely trustworthy agent (and corresponds to zero utility loss).[22]

In any case, we can now define the trust $T(\beta,X(O))$ of an agent α is an agent β over a particular set $X(O) = \{x_1, \ldots, x_k\}$ of issues appearing in the contract O (or in the expanded one O_+ as an aggregation of the trust in each individual issue (e.g., trust in delivering on time, paying on time, and the product having the quality specified in the contract). That is, we postulate $T(\beta, X(O) = agg(T(\beta x_1), \ldots, T(\beta,x_k))$, where agg: $[0, 1]^k \rightarrow [0, 1]$ is a *suitable* aggregation function.[23] If some issues are considered to be more important than others, the aggregation function should take this into consideration. This can be achieved by means of different weights given for each issue $x_i \in X(O)$ (the higher the weight, the more important the issue). A typical choice would be to take the aggregation[24] function as a weighted mean:

$$T(\beta, X(O)) = \sum_{x_i \in X(O)} w_i \cdot T(\beta, x_i) \tag{3}$$

where $\Sigma w_i = 1$ and $0 \leq w_i \leq 1$.

TRUST IN PRACTICE

Trust models are not useful in their own right. Rather they need to be coupled with an agent's decision model such that the agent is able to perform some of its tasks better. In our model, when an agent, say α, has a particular task to contract for, it will decide on the issues to be negotiated and identify possible interaction partners, say $\{\beta_1, \beta 2, \ldots, \beta_p\} \subseteq Ag$. For each agent in this set, we can calculate the trust value for each issue (i.e., $T(\beta, x)$) and aggregate these to give a general trust value for each agent (using Eq. 3). That is, $T(\beta_1, X'), T(\beta_2, X'), \ldots, T(\beta_p, X')$, where $X' \subseteq X$ is the set of issues under consideration. Trust can thus provide an ordering of the agents in terms of their overall reliability for a proposed contract. Agent α can then easily choose the preferred agent or the set of agents it would want to negotiate with (i.e., by choosing the most trustworthy one(s)).

In addition, however, we use trust to influence the negotiation that takes place before an agreement is signed. In this way, trust can directly influence the quality of agreements reached and the efficiency of the negotiation. First, the expected range used to obtain the trust value is instead used to change the range of negotiable values of a particular issue. Second, trust is used to change the number of issues to be negotiated.

Re-defining Negotiation Intervals

At contracting time, issue-value assignments, $x = v$, are agreed upon. Agents accept values that lie within a range $[v_{min}, v_{max}]$, such that $U_x^{\alpha}(v_{min}) > 0$ and $U_x^{\alpha}(v_{max}) > 0$. This interval is the acceptable range that an agent uses to offer and counter offer (according to a strategy) during negotiation (Jennings et al. 2001). Moreover, given a potential issue-value assignment $x = v_0$ in an offer, an agent can compute an interval of expected values. Thus, using Eq. 2, we have $EV_{cr}(\beta, x, v_0) = [ev^-, ev^+]$ over which the value v' actually obtained after execution is likely to vary. This range defines the uncertainty in the value of the issue and if the acceptable range $[ev^-, ev^+]$ does not fit within $[v_{min}, v_{max}]$, there exists the possibility that the final value may lie outside the acceptable region. This, in turn, means that $U_x^{\alpha}(v')$ may be zero, which is clearly undesirable and irrational.

Given this information, the agent can strategically restrict the negotiation interval $[v_{min}, v_{max}]$ with respect to the set of expected values $[ev^-, ev^+]$ as shown in the following. To do this, we first define the set of possible contracts, $\overline{\mathcal{O}}_x$, that are consistent with the expected values of x and its acceptance range, and then define the corrected values for v_{min} and v_{max}:

$$v'_{min} = \inf\{v | (x = v) \in O, O \in \overline{\mathcal{O}}_x\} \text{ and } v'_{max} = \sup\{v | (x = v) \in O, O \in \overline{\mathcal{O}}_x\}$$

where

$$\overline{\mathcal{O}} = \{O | (x = v) \in O, EV_{cr}(\beta, x, v) \subseteq [v_{min}, v_{max}]\}$$

This will shrink the range of negotiable values for an issue (i.e., from $[v_{min}, v_{max}]$ to $[v'_{min}, v'_{max}]$, where either $v'_{min} \geq v_{min}$ or $v'_{max} \leq v_{max}$, depending on which of the two limits v'_{min} and v'_{max} gives higher utility respectively) to ensure that the final outcome will fit within the range $[v_{min}, v_{max}]$. As well as reducing the possibility that the executed value will lie outside the acceptable range, reducing the negotiation range can also bring some other added benefits. It can help the agent reduce the time to negotiate over the value of each issue (e.g., if the range is smaller, the number of possible offers is also smaller) and it can help the agent to make better decisions that depend on the negotiation outcome (e.g., if a seller is expected to deliver goods one

day later than the agreed three days, the buyer can adjust its other tasks to fit with delivery in four days).

Extending the Set of Negotiable Issues

Initially, we argued that higher trust could reduce the negotiation dialogue and lower trust could increase the number of issues negotiated over. In this section, we deal with this particular use of trust in defining the issues that need to be negotiated. To this end, issues not explicitly included in a contract O^{α} may receive an expected value through one of the rules in *Rules*(α) for an agent α: **r: If** $x_1 = \upsilon_1$ *and* $x_2 = \upsilon_2$ *and* ... $x_m = \upsilon_m$ **Then** $x = \upsilon$.

Thus, if the premise of such a rule is true in a contract, the issue x is expected to have the value υ. If, however, the trust in the agent fulfilling the values of the issues present in the premises is not very high, it means that the agent believes that the values $\upsilon_1, \upsilon_2, \ldots, \upsilon_n$ may not be eventually satisfied. In such a case, to ensure that the issue x actually receives value υ, it should be added to the negotiated terms of the contract. This means that, when the trust is low in the premises, the unspecified issues (as discussed earlier) are added to the contracted issues in order to try and ensure that they will be met (whereas if trsut is high the issue is not negotiated). For example, if a buyer believes that the quality of a product to be delivered (the premise of a rule) will not be the quality of the product actually delivered, the buyer might request that the product satisfies very specific standards (e.g., kitemark or CE), which it privately expected and would not normally specify in a contract if trust were high.

Formally, this means that if $T(\beta, X_{\mathbf{r}}) \leq threshold$, (where ($T(\beta, X_{\mathbf{r}})$ is defined as per Eq. 3 and $X_{\mathbf{r}}$ is the set of issues in the premise of rule **r**), then the issue x in the conclusion of the rule should be added to the set of contract terms. On the other hand, as an agent becomes more confident that its interaction partner is actually performing well on the issues in the contract, it might eventually be pointless negotiating on the issue if the premises of the issue presuppose that the value expected will actually be obtained. This is, if $T(\beta, X_{\mathbf{r}}) > threshold$, then the issue x in the conclusion of the rule can be removed from the set of contract terms.

The two processes described here serve to expand and shrink the space of negotiation issues. For a new entrant to the system, for example, the trust value others have in it are likely to be low and hence the number of issues negotiated over will be large. But, as it acquires the trust of others, the number of issues it would need to negotiate will go down. Ultimately, with more trust, the set of negotiable issues can thus be reduced to a minimal set, affording shorter negotiation dialogues. Conversely, with less trust, the negotiable issues expand, trading off the length of dialogues with higher expected utility.

CONCLUSIONS AND FUTURE WORK

In this paper, we have discussed the necessary components to build up a concrete computational trust model based on direct and indirect multi-agent interactions. We have instantiated context, risk, and confidence values using rules that apply over the issues negotiated in a contract. From these components a measure of trust has been developed. Moreover, we have shown the worth of our model by describing how it can directly guide an agent's decisions in (i) choosing interaction partners, (ii) devising the set of negotiable issues, and (iii) determining negotiation intervals. The latter enable an agent to reduce time spent in negotiation dialogues, choose more reliable contractors, and adapt the negotiation strategy to each contractor. These are not possible using current trust models.

Future work will focus on studying and refining the properties of the model for both cooperative and competitive settings through simulations. Also, we aim to investigate modifications to the model to take into account relative utility variations. Finally, the trust measure will be made more sensitive to the stance taken by an opponent during the negotiation dialogue (e.g., if the opponent provided arguments backing its reliability).

NOTES

1. Risk is here defined as the maximum (utility) loss an agent can expect given an opponent reneges on its commitments in a given interaction (Marsh 1994; Zeckhauser and Viscusi 1990).
2. We build upon our work in Ramchurn et al. (2003).
3. If $\mathcal{G}$ denotes a partition $\{G_1, G_2, \ldots, G_i\}$ of the society of agents into non-empty groups, then for all $G_i, G_j, \in \mathcal{G}, G_i \cup G_j = \emptyset, \cap_i G_i = A_g$.
4. We believe these are the necessary, rather than sufficient, sets of norms that can give rise to unspecified expectations. Other sets of norms could arise from agents creating them or from legal systems, for example.
5. Norms can be of a very complex nature. However, in this paper, we *operationalize* norms in the form of constraints that apply over the values of issues in a contract and foresee using richer representations of norms in future work.
6. Richer syntaxes could also be thought of for premises in these rules, allowing for predicates like $\geq, \leq, \neq$, but equality $(=)$ will suffice for the purposes of this paper.
7. Again, we consider these features as necessary rather than sufficient. More features could be added (e.g., social relationships existing between agents or reasons given by an agent explaining its poor performance) and their impact will be investigated in future work.

8. We do not specify the institutional rules as part of the context since the decision to choose an institution is not defined by the context. However, these rules need to be specified before an agent is able to calculate its trust in its opponent. Depending on the level of trust, the opponent may then be chosen as an interaction partner (as will be shown later).
9. By taking into account such a dynamic context in evaluating trust, the model aims to adapt to cases where the environment and the agent are not necessarily static.
10. Our definition of *confidence* generally caters to a variety techniques that could be used to derive confidence values. In future work, we aim to define more explicit semantics of confidence values and enrich our definition of confidence.
11. The model can be made to work on relative deviations by making simple modifications to the fuzzy sets and the technique used to elicit confidence levels. This will be investigated in future work.
12. We also assume that the translation between the common and the specific terms is private. However, we do require that the common terms have the same agreed upon interpretations among the agents in order to permit a meaningful communication of reputation values.
13. The shape of the membership function given only serves as an example. Arbitrarily complex functions can be used in reality.
14. There are several techniques to model this range using probability distributions given the size of samples of ΔU_x that can be obtained from the interaction history (e.g., binomial, normal, or poisson distributions). Moreover, the behavior of the agent could also be modeled as a time-series so as to predict its behavior over future time points or analyzed using other data-mining techniques (e.g., cluster analysis, neural networks). However, the more complex the analysis, the more time and memory the algorithm will need to devise a level of confidence.
15. The type of probability distribution is not central to the trust model we wish to devise, provided it is continuous, and there are techniques to estimate the mean and variance given a small sample of values (since the agent's interactions will certainly not generate the infinite number of samples/points required to model a distribution accurately).
16. This assumes that the institution fully insures against any losses. This assumption could be removed and a risk level determined according to the institutional rules as well.
17. We are therefore implicitly assuming that all these measures are commensurate (i.e., have the same meaning and are based on the same scale), and hence their aggregation make sense.
18. We could also use other reputation measures provided by REGRET, such as neighborhood or system reputation. We intend to explore these measures and others in future work.

19. It is important to specify that only those completed interactions should be taken into account since only these can give us information about the behavior of the opponent in its execution of contracts. Negotiations could end up in no agreements and these should be excluded when counting interactions in the case base.
20. By using utility variations, rather than value variations, we can use the same membership functions even if the utility function changes over time.
21. We acknowledge that other bounds may be applied in other trust models (e.g., [−1,1] as in Marsh[1994] or [0, ∞] in eBay). See Marsh (1994) for a wider discussion on the meaning of the bounds on the rating.
22. Our choice for the bounds of [0, 1] serves to simplify the analysis when normalizing all trust ratings in issues and over contracts.
23. Generally, an aggregation function is monotonic such that $\min(u_1, \ldots, u_k) \leq g(u_1, \ldots, u_k) \leq \max(u_1, \ldots, u_k)$ (see [Calvo et al. 2002] for a survey).
24. While most aggregation operators are defined parametrically with respect to weights assigned to each component to be aggregated, more sophisticated aggregation models (based, for example, on different Lebesgue, Choquet, or Sugeno integrals) could also be used (Calvo et al. 2002).

REFERENCES

Calvo, T., A. Kolesarova, M. Komornikova, and R. Mesiar. 2002. Aggregation operators: Properties, classes and construction methods. In *Studies in Fuzziness and Soft Computing*, eds. T. Calvo, G. Mayor, and R. Mesiar, Vol 97, 3–104. Springer.

Dasgupta, P. Trust as a commodity. In *Trust: Making and Breaking Cooperative Relations*, ed. D. Gambetta, 49–72. Oxford: Blackwell.

Esteva, M., J. A. Rodríguez, C. Sierra, P. Garcia, and J. L. Arcos. 2001. On the formal specifications of electronic institutions. *Lecture Notes in Artificial Intelligence* 1991:126–147.

Jennings, N. R., P. Faratin, A. R. Lomuscio, S. Parsons, C. Sierra, and M. Wooldridge. 2001. Automated negotiation: prospects, methods and challenges. *International Journal of Group Decision and Negotiation* 10(2):199–215.

Marsh, S. 1994. *Formalising Trust as a Computational Concept*. Ph.D. thesis, Department of Mathematics and Computer Science, University of Stirling.

Molm, L. D., N. Takahashi, and G. Peterson. 2000. Risk and trust in social exchange: An experimental test of a classical proposition. *American Journal of Sociology* 105:1396–1427.

Ramchurn, S. D., D. Huynh, and N. R. Jennings. 2004. Trust in multi-agent systems. *The Knowledge Engineering Review* 19(1).

Ramchurn, S. D., C. Sierra, L. Godo, and N. R. Jennings. 2003. A computational trust model for multi-agent interactions based on confidence and reputation. In *Workshop on Deception, Trust, and Fraud, AAMAS*, eds. R. Falcone, S. Barber, L. Korba, and M. Singh, 69–75.

Resnick, P. and R. Zeckhauser. 2002. Trust among strangers in Internet transactions: Empirical analysis of ebay's reputation system. In *Advances in Applied Microeconomics*, ed. M. R. Baye, volume 11, 127–157, Amsterdam: Elsevier Science.

Sabater, J. and C. Sierra. 2002. REGRET: a reputation model for gregarious societies. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*, eds. C. Castelfranchi, and L. Johnson, pages 475–482.

Yamagishi, T., K. Cook, and M. Watabe. 1998. Uncertainty, trust, and commitment formation in the United States and Japan. *American Journal of Sociology* 104:165–94.

Yu, B. and M. P. Singh. 2002. Distributed reputation management for electronic commerce. *Computational Intelligence* 18(4):535–549.

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

Zeckhauser, R. and W. K. Viscusi. 1990. Risk within reason. *Science* 248(4955):559–564.