Robust Trust: Prior Knowledge, Time and Context

John Debenham[1] and Carles Sierra[2]

[1] QCIS, University of Technology, Sydney, Australia
debenham@it.uts.edu.au
[2] Institut d'Investigació en Intel·ligència Artificial - IIIA,
Spanish Scientific Research Council, CSIC
08193 Bellaterra, Catalonia, Spain
sierra@iiia.csic.es

Abstract. *Trust* is an agent's expectation of the value it will observe when it evaluates the enactment of another agent's commitment. There are two steps involved in trust: first the action that another agent is expected to enact given that it has made a commitment, and second the expected valuation of that action when the result of that action is eventually consumed. A computational model of trust is presented that takes account of: *prior knowledge* of other agents, the evolution of trust estimates in *time*, and the evolution of trust estimates in response to changes in *context*. This model is founded on the principle of *information-based* agency that each and every utterance made contains valuable information. The computational basis for the model is substantially simpler and is more theoretically grounded than previously reported.

1 Introduction

The social concept of trust has received considerable attention. The seminal paper [1] describes two approaches to trust: first, as a belief that another agent will do what it says it will, or will reciprocate for common good, and second, as constraints on the behaviour of agents to conform to trustworthy behaviour. This paper is concerned with the first approach where trust is something that is learned and evolves. [2] presents a comprehensive categorisation of trust research: policy-based, reputation-based, general *and* trust in information resources. [3] presents an interesting taxonomy of trust models in terms of nine types of trust model. The scope described there fits well with this work with the possible exception of identity trust and security trust that are out of scope. [4] describes a powerful model that integrates interaction and role-based trust with witness and certified reputation that also relate closely to our model. *Reputation* is the opinion (more technically, a social evaluation) of a group about something — in a social environment — reputation feeds into trust [5].

The informal meaning of the statement "agent α trusts agent β" is that α expects β to act in a way that is somehow preferred by α. Human agents seldom trust another for *any* action that they may take — it is more usual to develop

C. Huemer and P. Lops (Eds.): EC-Web 2012, LNBIP 123, pp. 1–12, 2012.

Action	Sign	Enact	Evaluate
Object	*(a,b)*	*(a',b')*	*b'*
Time	*t*	*t'*	*t''*

Fig. 1. Contract signing, execution and evaluation

a trusted expectation with respect to a particular set of actions. For example, "I trust John to deliver fresh vegetables" whilst the trustworthiness of John's advice on investments may be terrible. In this paper we discuss trust when the set of actions is restricted to negotiating, signing and enacting contracts that are expressed using some particular ontology. This then excludes morally founded trust as in: "Can I trust you to *do the right thing*?".

We assume that a multiagent system, $\{\alpha, \beta_1, \ldots, \beta_o, \xi, \theta_1, \ldots, \theta_t\}$, containing an agent α that interacts with negotiating agents, $\mathcal{X} = \{\beta_i\}$, information providing agents, $\mathcal{I} = \{\theta_j\}$, and an *institutional agent*, ξ, that represents the institution where we assume the interactions happen. Institutions give a normative context to interactions that simplify matters (e.g an agent can't make an offer, have it accepted, and then renege on it). The institutional agent ξ may form opinions on the actors and activities in the institution and may publish reputation estimates on behalf of the institution. The agent ξ also fulfils a vital role to compensate for any lack of sensory ability in the other agents by promptly and accurately reporting observations as events occur. For example, without such reporting an agent may have no way of knowing whether it is a fine day or not.

Our agents are information-based [6], they are endowed with machinery for valuing the information that they have, and that they receive. Information-based agency was inspired by the observation that "everything an agent says gives away information". They model how much they know about other agents, and how much they believe other agents know about them. Everything in their world, including their information, is uncertain; their only means of reducing uncertainty is acquiring fresh information. To model this uncertainty, their world model, $\mathcal{M}^t$, consists of random variables each representing a point of interest in the world. Distributions are then derived for these variables on the basis of information received. Over time agents acquire large amounts of information that are distilled into convenient measures including trust. By classifying private information into functional classes, and by drawing on the structure of the ontology, information-based agents develop other measures including a map of the 'intimacy' [7] of their relationships with other agents.

Section 2 discusses the notion of trust and develops a formal characterisation of it. The core mechanism for maintaining trust estimates is described in Section 3. Prior knowledge is then taken into account in Section 4 that includes a discussion of the *reliability* of an agent's utterances. Time is discussed in Section 5 and Context in Section 6. Section 7 concludes with a discussion of future work.

2 The Notion of 'Trust'

In this paper trust is concerned with valuing enactments made in fulfilment of commitments expressed in contracts. The scenario is: two agents α and β negotiate with the intention of leading to a *signed contract* that is a pair of commitments, (a, b), where a is α's and b is β's. A contract is signed by both agents at some particular time t. At some later time, t', both agents will have enacted their commitments[1] in some way, as say (a', b'). At some later time again, t'', α will consume b' and will then be in a position to evaluate the extent to which β's enactment of (a, b), b', was in α's interests. See Figure 1.

α's trust of agent β is expressed as α's expectation of its eventual valuation of β's future actions. We consider how α forms these expectations, how α will compare those expectations with observations, and how α then determines whether β's actions are preferred to α's expectations of them.

α forms expectations of β's future actions on the basis of all that it has: its full interaction history $H_\alpha \in \mathcal{H}_\alpha$ where $\mathcal{H}_\alpha$ is the set of all possible interaction histories that may be expressed in α's ontology[2]. H_α is a record of all interactions with each negotiating agent in $\mathcal{X}$ and with each information providing agent in $\mathcal{I}$. Let $\mathcal{B} = (b_1, b_2, \dots)$ denote that space of all enactments that β may make and $\mathcal{A}$ the space of α's enactments. Assuming that the space of contracts and enactments are the same, the space of all contracts and enactments is: $\mathcal{C} = \mathcal{A} \times \mathcal{B}$.

This raises the strategic question of given an expectation of some particular future requirements how should α strategically shape its interaction history to enable it to build a reliable expectation of β's future actions concerning the satisfaction of those particular requirements [8]. At time t'' α compares b' with α's expectations of β's actions, β having committed at time t to enact b at time t'. That is:

$$\text{compare}_\alpha^{t''}(\mathbb{E}_\alpha^t(\text{Enact}_\beta^{t'}(b)|\text{sign}_{\alpha,\beta}^t((a,b)), H_\alpha^t), b')$$

where $\text{sign}_{\alpha,\beta}^t((a,b))$ is a predicate meaning that the joint action by α and β of signing the contract (a, b) was performed at time t, and $\text{Enact}_\beta^{t'}(b)$ is a random variable over $\mathcal{B}$ representing α's expectations over β's enactment action at time t', $\mathbb{E}_\alpha^t(\cdot)$ is α's expectation, and $\text{compare}(\cdot, \cdot)$ somehow describes the result of the comparison.

Trust is the expectation of *the evaluation of* β's enactments made in fulfilment of its contractual commitments. Let $\mathcal{V} = (v_1, v_2, \dots, v_V)$ be the valuation space then α's expectation of the evaluation of a particular action that β may make is represented as a probability distribution over $\mathcal{V}$: $(f_1, f_2, \dots, f_V)$. We expect the set $\mathcal{V}$ to be smaller than the set $\mathcal{B}$, and so developing a sense of expectation for the value of β's actions should be easier than for the actions themselves. That is, we consider the expectation:

$$\mathbb{E}_\alpha^t(\text{Value}_\beta^{t''}(b)|\text{sign}_{\alpha,\beta}^t((a,b)), H_\alpha^t)$$

[1] For convenience we assume that both agents are presumed to have been completed their enactments by the same time, t'.

[2] The ontology is not made explicit to avoid overburdening the discussion.

where $\text{Value}^{t''}(b)$ is a random variable over $\mathcal{V}$ representing α's expectations of the value of β's enactment action given that he signed (a, b) and given H^t_α. At time t'' it then remains to compare expectation, $\mathbb{E}^t_\alpha(\text{Value}^{t''}_\beta(b)|\text{sign}^t_{\alpha,\beta}((a,b)), H^t_\alpha)$, with observation, $\text{val}_\alpha(b')$, where $\text{val}(\cdot)$ represents α's preferences — i.e. it is α's utility function[3].

We are now in a position to define 'trust'. *Trust*, $\tau_{\alpha\beta}(b)$, is a computable[4] [9] estimate of the distribution: $\mathbb{E}^t_\alpha(\text{Value}^{t''}_\beta(b)|\text{sign}^t_{\alpha,\beta}((a,b)), H^t_\alpha)$. τ is a summarising function that distils the trust-related aspects of the (probably very large) set H_α into a probability distribution that may be computed. $\tau_{\alpha\beta}(b)$ summarises the large set H_α. The set of contracts $\mathcal{C}$ is also large. It is practically unfeasible to estimate trust for each individual contract. To deal with this problem we appeal to the structure of the ontology, and aggregate estimates into suitable classes such as John's trustworthiness in supplying Australian red wine.

In real world situations the interaction history may not reliably predict future action, in which case the notion of trust is fragile. No mater how trust is defined we expect trusted relationships to develop slowly over time. On the other hand they can be destroyed quickly by an agent whose actions unexpectedly fall below expectation. This highlights the importance of being able to foreshadow the possibility of untrustworthy behaviour.

$\tau_{\alpha\beta}(b)$ is predicated on α's ability to form an expectation of the value of β's future actions. This is related to the famous question posed by Laplace "what is the probability that the sun will rise tomorrow?". Assuming that it has always previously been observed to do so and that there have been n prior observations then if the observer is in complete ignorance of the process he will assume that the probability distribution of a random variable representing the prior probability that the sun will rise tomorrow is the maximum entropy, uniform distribution on $[0, 1]$, and using Bayes' theorem will derive the posterior estimate $\frac{n+1}{n+2}$. The key assumption being that the observer is "in complete ignorance of the process". There may be many reasons why the sun may not rise such as the existence of a large comet on a collision trajectory with earth. These all important reasons are the *context* of the problem.

Laplace's naïve analysis above forms the basis of a very crude measure of trust. Suppose that the valuation space is: $\mathcal{V} = \{\text{bad}, \text{good}\}$, and that α is considering signing contract (a, b) with β. Let the random variable B denote the value of β's next action. Then assuming that we know nothing about the contract or about β except that this contract has been enacted by β on n prior occasions and that the valuation was "good" on s of those occasions. Using the maximum entropy prior distribution for B, $[0.5, 0.5]$, Bayes' theorem gives us a posterior distribution $[\frac{n-s+1}{n+2}, \frac{s+1}{n+2}]$. If at time t α signs the contract under consideration then the expected probability of a "good" valuation at time t'' is:

[3] It is arguably more correct to consider: $\text{Value}((a, b)) = \text{Value}(b) - \text{Value}(a)$, as β's actions may be influenced by his expectations of α's enactment of a — we choose to avoid this additional complication.

[4] *Computable* in the sense that it is easy to compute and not simply Turing computable.

$\frac{s+1}{n+2}$. This crude measure has little practical value although it readily extends to general discrete valuation spaces, and to continuous valuation spaces. The zero-information, maximum entropy distribution is the *trivial trust measure*. The crude Laplacian trust measure is in a sense the simplest non-trivial measure.

The weaknesses of the crude Laplacian trust measure above show the way to building a reliable measure of trust [10]. These are:

Prior knowledge. The use of the maximum entropy prior[5] is justified when there is absolutely no prior knowledge or belief of an agent's behaviour. In practical scenarios we expect prior observations, reputation measures or the opinions of other agents to be available to be reflected in the prior.

Time. There is no representation of time. In the crude trust measure all prior observations have the same significance, and so an agent that used to perform well and is deteriorating may have the same trust measure as one that used to perform badly and is now performing well.

Context. There is no model of general events in the world or of *how* those events may effect an agent's behaviour. This includes modelling causality, *why* an agent might behave as it does.

This section defines trust as a historic[6] estimator of the expected value of future enactments, and concluded with three features of a reliable measure of trust. This section also described the fundamental role that the structure of the ontology plays in the trust model. Following sections describe such a measure that uses new and improved computational methods of information-based agents [6] particularly their information evaluation, acquisition and revelation strategies that ideally suits them to this purpose. The core trust mechanism is detailed in Section 3 and subsequent sections then detail the incorporation of prior knowledge, time and context.

3 The Core Mechanism

Section 2 ends with three essential components of a reliable trust model. Those three components will be dealt with in due course. In this section we describe the core trust estimation mechanism. In subsequent sections we enhance the core with the three essential components. The final component, context, is incomplete as it relies on the solution to unsolved problems that are beyond the scope of this paper.

The general idea is that trust estimates are updated whenever α evaluates $\text{val}_{\alpha}^{t''}(b')$ for some previously signed contract (a, b). The contract space is typically very large and so estimates are not maintained for individual contracts; instead they are maintained for selected abstractions based on the ontology. Abstractions

[5] The maximum entropy prior expresses total uncertainty about what the prior distribution is.

[6] *Historic* in the sense that the estimation can be performed on the basis of the agent's interaction history.

are denoted by the 'hat' symbol: e.g. $\hat{a}$. For example, "red wine orders for more that 24 bottles" or "supply of locally produced cheese". As we will see when an evaluation $\text{val}_{\alpha}^{t''}(b')$ is performed, the trust estimates, $\tau_{\alpha\beta}(\hat{b})$, for certain selected nearby abstractions, $\hat{b}$, are updated.

In the absence of incoming information the integrity of an information-based agent's beliefs decays in time. In the case of the agent's beliefs concerning trust, incoming information is in the form of valuation observations $\text{val}_{\alpha}^{t''}(b')$ for each enacted contract. If there are no such observations in an area of the ontology then the integrity of the estimate for that area should decay.

In the absence of valuation observations in the region of $\hat{b}$, $\tau_{\alpha\beta}(\hat{b})$ decays to a *decay limit distribution* $\overline{\tau_{\alpha\beta}(\hat{b})}$ (denoted throughout this paper by 'overline'). The decay limit distribution is the zero-data distribution, but not the zero-information distribution because it takes account of reputation estimates and the opinions of other agents [11]. We assume that the decay limit distribution is known for each abstraction $\hat{b}$. At time s, given a distribution for random variable $\tau_{\alpha\beta}(\hat{b})^s$, and a decay limit distribution, $\overline{\tau_{\alpha\beta}(\hat{b})^s}$, $\tau_{\alpha\beta}(\hat{b})$ decays by:

$$\tau_{\alpha\beta}(\hat{b})^{s+1} = \Delta(\overline{\tau_{\alpha\beta}(\hat{b})^s}, \tau_{\alpha\beta}(\hat{b})^s) \tag{1}$$

where s is time and Δ is the *decay function* for the X satisfying the property that $\lim_{s\to\infty} \tau_{\alpha\beta}(\hat{b})^s = \overline{\tau_{\alpha\beta}(\hat{b})}$. For example, Δ could be linear:

$$\tau_{\alpha\beta}(\hat{b})^{s+1} = (1-\mu) \times \overline{\tau_{\alpha\beta}(\hat{b})^s} + \mu \times \tau_{\alpha\beta}(\hat{b})^s$$

where $0 < \mu < 1$ is the decay rate.

We now consider what happens when valuation observations are made. Suppose that at time s, α evaluates β's enactment b' of commitment b, $\text{val}_{\alpha}^{s}(b') = v_k \in \mathcal{V}$. The update procedure updates the probability distributions for $\tau_{\alpha\beta}(\hat{b})^s$ for each $\hat{b}$ that is "moderately close to" b. Given such a $\hat{b}$, let $\mathbb{P}^s(\tau_{\alpha\beta}(\hat{b}) = v_k)$ denote the prior probability that v_k would be observed. The update procedure is in two steps. First, we estimate the posterior probability that v_k would be observed, $\mathbb{P}^{s+1}(\tau_{\alpha\beta}(\hat{b}) = v_k)$ for the particular value v_k. Second, we update the entire posterior distribution for $\tau_{\alpha\beta}(\hat{b})$ to accommodate this revised value.

Given a $\hat{b}$, to revise the probability that v_k would be observed we work with three things: the observation: $\text{val}_{\alpha}^{s}(b')$, the prior: $\mathbb{P}^s(\tau_{\alpha\beta}(\hat{b}) = v_k)$, and the decay limit value: $\mathbb{P}^s(\overline{\tau_{\alpha\beta}(\hat{b})} = v_k)$. The observation $\text{val}_{\alpha}^{s}(b')$ may be represented as a probability distribution with a '1' in the k'th place and zero elsewhere, $\boldsymbol{u}_k$. To combine it with the prior we discount its significance for two reasons:

- b may not be semantically close to $\hat{b}$, and
- $\text{val}_{\alpha}^{s}(b') = v_k$ is a single observation whereas the prior distribution represents the accumulated history of previous observations.

To discount the significance of the observation $\text{val}_{\alpha}^{s}(b') = v_k$ we determine a value in the range between '1' and the zero-data, decay limit value $\mathbb{P}^s(\overline{\tau_{\alpha\beta}(\hat{b})} = v_k)$ by:

$$\delta = \mathrm{Sim}(b, \hat{b}) \times \kappa + (1 - \mathrm{Sim}(b, \hat{b}) \times \kappa) \times \mathbb{P}^s(\overline{\tau_{\alpha\beta}(\hat{b})} = v_k) \tag{2}$$

where $0 < \kappa < 1$ is the learning rate, and $\mathrm{Sim}(\cdot,\cdot)$ is a semantic similarity function. Then the posterior estimate $\mathbb{P}^{s+1}(\tau_{\alpha\beta}(\hat{b}) = v_k)$ is given by:

$$\mathbb{P}^{s+1}(\tau_{\alpha\beta}(\hat{b}) = v_k) = \frac{\rho\delta(1-\omega)}{\rho\delta(1-\omega) + (1-\rho)(1-\delta)\omega} = \nu \tag{3}$$

where δ is given by Equation 2, $\rho = \mathbb{P}^s(\tau_{\alpha\beta}(\hat{b}) = v_k)$ is the prior value, and $\omega = \mathbb{P}^s(\overline{\tau_{\alpha\beta}(\hat{b})} = v_k)$ is the decay limit value. That is, we combine the two 'observed' probabilities ρ and δ in the context of the pre-observation value ω.

It remains to update the entire posterior distribution for $\tau_{\alpha\beta}(\hat{b})$ to accommodate the constraint $\mathbb{P}^{s+1}(\tau_{\alpha\beta}(\hat{b}) = v_k) = \nu$. Information-based agents [6] employ a standard procedure for updating distributions, $\mathbb{P}^t(X = x)$ subject to a set of linear constraints on X, $c(X)$, using:

$$\mathbb{P}^{t+1}(X = x | c(X)) = \mathrm{MRE}(\mathbb{P}^t(X = x), c(X))$$

where the function MRE is defined by: $\mathrm{MRE}(\boldsymbol{q}, \boldsymbol{g}) = \arg\min_{\boldsymbol{r}} \sum_j r_j \log \frac{r_j}{q_j}$ such that $\boldsymbol{r}$ satisfies $\boldsymbol{g}$, $\boldsymbol{q}$ is a probability distribution, and $\boldsymbol{g}$ is a set of n linear constraints $\boldsymbol{g} = \{g_j(\boldsymbol{p}) = \boldsymbol{a_j} \cdot \boldsymbol{p} - c_j = 0\}, j = 1, \ldots, n$ (including the constraint $\sum_i p_i - 1 = 0$). The resulting $\boldsymbol{r}$ is the *minimum relative entropy distribution*[7] [12]. Applying this procedure to $\tau_{\alpha\beta}(\hat{b})$:

$$\mathbb{P}^{s+1}(\tau_{\alpha\beta}(\hat{b}) = v) = \mathrm{MRE}(\mathbb{P}^s(\tau_{\alpha\beta}(\hat{b}) = v), \mathbb{P}^{s+1}(\tau_{\alpha\beta}(\hat{b}) = v_k) = \nu)$$

where ν is the value given by Equation 3.

Whenever α evaluates an enactment $\mathrm{val}^s_\alpha(b')$ of some commitment b, the above procedure is applied to update the distributions for $\mathbb{P}(\tau_{\alpha\beta}(\hat{b}) = v)$. It makes sense to limit the use of this procedure to those distributions for which $\mathrm{Sim}(b, \hat{b}) > y$ for some threshold value y.

4 Prior Knowledge

The decay-limit distribution plays a key role in the estimation of trust. It is not directly based on any observations and in that sense it is a "zero data" trust estimate. It is however not "zero information" as it takes account of opinions and reputations communicated by other agents [11]. The starting point for constructing the decay-limit distribution is the maximum entropy (zero-data, zero-information) distribution. This gives a two layer structure to the estimation of trust: opinions and reputations shape the decay-limit distribution that in turn

[7] This may be calculated by introducing Lagrange multipliers $\boldsymbol{\lambda}$: $L(\boldsymbol{p}, \boldsymbol{\lambda}) = \sum_j p_j \log \frac{p_j}{q_j} + \boldsymbol{\lambda} \cdot \boldsymbol{g}$. Minimising L, $\{\frac{\partial L}{\partial \lambda_j} = g_j(\boldsymbol{p}) = 0\}, j = 1, \ldots, n$ is the set of given constraints $\boldsymbol{g}$, and a solution to $\frac{\partial L}{\partial p_i} = 0, i = 1, \ldots, I$ leads eventually to $\boldsymbol{p}$.

plays a role in forming the trust estimate that takes account of observed data [13]. Communications from other agents may not be reliable. α needs a means of estimating the reliability of other agents before they can be incorporated into the decay-limit distribution — reliability is discussed at the end of this section.

Reputation is the opinion (more technically, a social evaluation) of a group about something. So a group's reputation about a thing will be related *in some way* to the opinions that the individual group members hold towards that thing, or to shared evaluations that they may hold. An *opinion* is an assessment, judgement or evaluation of something. Opinions are represented in this paper as probability distributions on a suitable ontology that for convenience we identify with the *evaluation space* $\mathcal{V}$. That is, we assume that opinions communicated by β concerning another agent's trustworthiness are expressed as predicates using the same valuation space as $\mathcal{V}$ over which α represents its trust estimates.

An opinion is an evaluation of an *aspect* of a thing. A rainy day may be evaluated as being "bad" from the aspect of being suitable for a picnic, and "good" from the aspect of watering the plants in the garden. An aspect is the "point of view" that an agent has when forming his opinion. An opinion is evaluated in context. The *context* is everything that the thing is being, explicitly or implicitly, evaluated with or against. The set of valuations of all things in the context calibrates the valuation space. For example, "this is the best paper in the conference". The context can be vague: "of all the presents you could have given me, this is the best". If agents are to discuss opinions then they must have some understanding of each other's context.

Summarising the above, an *opinion* is an agent's evaluation of a particular aspect of a thing in context. A representation of an opinion will contain: the thing, its aspect, its context, and a distribution on $\mathcal{V}$ representing the evaluation of the thing. α acquires opinions and reputations through communication with other agents. α estimates the reliability of those communicating agents before incorporating that information into the decay-limit distributions. The basic process is the same for opinions and reputations; the following sub-section 4.1 describes the incorporation of opinions only.

4.1 The Decay-Limit Distribution and Reliability

Suppose agent β' informs agent α of his opinion of the trustworthiness of another agent β using an utterance of the form: $u = \texttt{inform}(\beta', \alpha, \tau_{\beta'\beta}(b))$, where conveniently b is in α's ontology. This information may not be useful to α for at least two reasons: β' may not be telling the truth, or β' may have a utility function that differs from α's. We will shortly estimate β''s "reliability", $R^t_\alpha(\beta')$ that measures the extent to which β' is telling the truth and that α and β' "are on the same page" or "think alike"[8]. Precisely, $0 < R^t_\alpha(\beta') < 1$; its value is used to moderate the effect of the utterance on α's decay-limit distributions. The estimation of $R^t_\alpha(\beta')$ is described below.

[8] The reliability estimate should perhaps also be a function of the commitment, $R^t_\alpha(\beta', b)$, but we choose to avoid that complication.

Suppose that α maintains the decay limit distribution $\overline{\tau_{\alpha\beta}(\hat{b})^s}$ for a chosen $\hat{b}$. In the absence of utterances informing opinions of trustworthiness, $\overline{\tau_{\alpha\beta}(\hat{b})^s}$ decays to the distribution with maximum entropy. As previously this decay could be linear:

$$\overline{\tau_{\alpha\beta}(\hat{b})^{s+1}} = (1-\mu) \times \text{MAX} + \mu \times \overline{\tau_{\alpha\beta}(\hat{b})^s}$$

where $\mu < 1$ is the decay rate, and MAX is the maximum entropy, uniform distribution.

When α receives an utterance of the form u above, the *decay limit distribution* is updated by:

$$\begin{aligned}&\overline{\tau_{\alpha\beta}(\hat{b})^{s+1}} \mid \texttt{inform}(\beta', \alpha, \tau_{\beta'\beta}(b)) = \\ &\quad\Big(1 - \kappa \times \text{Sim}(\hat{b}, b) \times R^s_\alpha(\beta')\Big) \times \overline{\tau_{\alpha\beta}(\hat{b})^s} + \kappa \times \text{Sim}(\hat{b}, b) \times R^s_\alpha(\beta') \times \tau_{\beta'\beta}(b)\end{aligned}$$

where $0 < \kappa < 1$ is the learning rate and $R^s_\alpha(\beta')$ is α estimate of β''s reliability. It remains to estimate $R^s_\alpha(\beta')$.

Estimating $R^s_\alpha(\beta')$ is complicated by its time dependency. First, in the absence of input of the form described following, $R^s_\alpha(\beta')$ decays to zero by: $R^{s+1}_\alpha(\beta') = \mu \times R^s_\alpha(\beta')$. Second, we describe how $R^s_\alpha(\beta')$ is increased by comparing the efficacy of $\overline{\tau_{\alpha\beta}(\hat{b})^s}$ and $\tau_{\beta'\beta}(b)^s$ in the following interaction scenario. Suppose at a time s, α is considering signing the contract (a, b) with β. α requests β''s opinion of β with respect to b, to which β may respond $\texttt{inform}(\beta', \alpha, \tau_{\beta'\beta}(b))$. α now has two estimates of β's trustworthiness: $\overline{\tau_{\alpha\beta}(\hat{b})^s}$ and $\tau_{\beta'\beta}(b)^s$. If α then signs the contract (a, b) at time t, and at some later time t'' evaluates β's enactment $\text{val}^{t''}_\alpha(b') = v_k$, say. $\overline{\tau_{\alpha\beta}(\hat{b})^s}$ and $\tau_{\beta'\beta}(b)^s$ are both probability distributions that each provide an estimate of $\mathbb{P}^s(\text{Value}_\beta(b) = v_k)$. If:

$$\mathbb{P}(\tau_{\beta'\beta}(b)^s = v_k) > \mathbb{P}(\overline{\tau_{\alpha\beta}(\hat{b})^s} = v_k)$$

then β''s estimate is better than α's and α increases $R^s_\alpha(\beta')$ using:

$$R^{s+1}_\alpha(\beta') = \kappa + (1-\kappa) \times R^s_\alpha(\beta')$$

where $0 < \kappa < 1$ is the learning rate.

5 Time

The core trust mechanism in Section 3 and the prior knowledge in Section 4 both give greater weight to recent observations than to historic. This may be a reasonable default assumption but has no general validity. Trust, $\tau_{\alpha\beta}(\hat{b})^s$, estimates *how* we expect β to act. If an agent is considering repeated interaction with β then he may also be interested in how β's actions are expected to *change* in time.

The way in which the trust estimate is evolving is significant in understanding which agents to interact with. For example, and agent for whom $\tau^s_{\alpha\beta}(\hat{b})$ is fairly constant in time may be of less interest than an agent who is slightly less trustworthy but whose trust is consistently improving. To capture this information we need something like the finite derivative: $\frac{\delta}{\delta s}\tau^s_{\alpha\beta}(\hat{b})$. The sum of the elements in such a vector will be zero, and in the absence of any data it will decay to the zero vector.

Estimating the rate of change of $\tau^s_{\alpha\beta}(\hat{b})$ is complicated by the way it evolves that combines continual integrity decay with periodic updates. Evolution due to decay tells us nothing about the rate of change of an agent's behaviour. Evolution caused by an update is performed following a period of prior decay, and may result in compensating for it. Further, update effects will be very slight in the case that the commitment b is semantically distant from $\hat{b}$. In other words, the evolution of $\tau^s_{\alpha\beta}(\hat{b})$ itself is not directly suited to capturing the rate of change of agent behaviour.

The idea for an indirect way to estimate how β's actions are evolving comes from the observation that $\tau_{\alpha\beta}(\hat{b})^s$ is influenced more strongly by more recent observations, and the extent to which this is so depends on the decay rate. For example, if the decay rate is zero then $\tau_{\alpha\beta}(\hat{b})^s$ is a time-weighted "average" of prior observations. Suppose that $\tau_{\alpha\beta}(\hat{b})^s$ has been evaluated. We perform a parallel evaluation using a lower decay rate to obtain $\tau^-_{\alpha\beta}(\hat{b})^s$, then the vector difference, $\tau_{\alpha\beta}(\hat{b})^s - \tau^-_{\alpha\beta}(\hat{b})^s$, is a vector the sum of whose elements is zero, and in which a positive element indicates a value that is presently "on the increase" compared to the historic average.

The preceding method for estimating change effectively does so by calculating a first difference. If we calculate another first difference using an even lower decay rate then we can calculate a second difference to estimate the *rate of* change. This may be stretching the idea too far!

6 Trust in Context

The informal meaning of context is information concerning everything in the environment that could effect decision making *together with* rules that link that information to the deliberative process. That is, *context* consists of facts about the environment *and* rules that link those facts to the agent's reasoning. Those rules may rely on common sense reasoning.

Human and artificial agents have rather different practical problems in dealing with context. One practical difficulty for human agents is assimilating new information in an information-overloaded environment. Humans then rely on common sense and experience to learn how to key contextual information to their deliberation. Storage permitting, artificial agents can assimilate real-time data flows with ease, and can manage the integrity decay of old information. After that things become tricky for artificial agents; identifying and dealing with inconsistency is a hard problem, and so is keying context to deliberation.

To make matters worse, both human and artificial agents can reasonably assume that their knowledge of their context is substantially incomplete. Dealing with context is arguably *the* major impediment to delivering trustworthy negotiation by artificial agents in the real world. After this grim observation we consider the context of trust.

Following the procedure described in Section 3 an agent builds up a sense of trust on the basis of its own past experience and statements of opinion and reputation from other agents. In a sense those statements of opinions and reputation are contextual information for the business of estimating trust.

Suppose that an agent has built up a sense of trust in another agent based on their prior interaction, before relying on that trust estimate as an indicator of future performance the agent will consider whether there are any perceivable changes in the context that cause it to distrust its previous observations as an indicator of future behaviour. As a simple example, if John has an impeccable history of delivering goods on time then the contextual information that John has sprained his ankle, or that he is overseas, may cause us to distrust our experience as an indicator of John's timeliness in the near future.

In this paper 'trust in context' is concerned with just one issue: is there any reason to distrust our trust estimate due to a change in context. Supposing that α is considering signing a contract (a, b) at time t, to address this issue we require:

1. knowledge of the context of previous observations of behaviour. Their *context* is the state of each of the observables in the environment and of the states of the other agents when those previous observations of behaviour were made — given the way that observations are aggregated in Section 3 the more recent the observation the greater its significance.
2. founded beliefs concerning the context that will pertain at the future time of the evaluation of the presumed future behaviour — i.e. at time t'' in Figure 1.
3. some reasoning apparatus that enables us to decide whether differences between the believed future context and the observed previous contexts cause us to modify our experience-based trust estimate.

Taken together these three points are the *context* of the trust estimate that α has for the act of signing (a, b) with β. As stated the context of an observation of behaviour is the state of *all* observables at the time the observation is made. This is a potentially massive exercise. A causal model that identified only those observables that could be seen to cause or affect the behaviour would simplify things but is a major issue in its own right and is beyond the scope of this discussion.

The information-based architecture makes a modest contribution to the maintenance of trust estimates through the persistent decay of information integrity by Equation 1. Beyond that we offer no 'magic bullet' solutions to the contextual problems described above and leave the discussion as a pointer to the work that is required to increase the reliability of trust estimation in dynamic environments.

7 Future Work

Current work is focussed heavily on the issues of context identified in Section 6. In particular we are exploring the application of the minimum message length principle to reduce the complexity of models of context — unfortunately this comes with a very high computational overhead.

References

1. Ramchurn, S., Huynh, T., Jennings, N.: Trust in multi-agent systems. The Knowledge Engineering Review 19(1), 1–25 (2004)
2. Artz, D., Gil, Y.: A survey of trust in computer science and the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 5(2), 58–71 (2007)
3. Viljanen, L.: Towards an Ontology of Trust. In: Katsikas, S.K., López, J., Pernul, G. (eds.) TrustBus 2005. LNCS, vol. 3592, pp. 175–184. Springer, Heidelberg (2005)
4. Huynh, T., Jennings, N., Shadbolt, N.: An integrated trust and reputation model for open multi-agent systems. Autonomous Agents and Multi-Agent Systems 13(2), 119–154 (2006)
5. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review 24(1), 33–60 (2005)
6. Sierra, C., Debenham, J.: Information-based agency. In: Proceedings of Twentieth International Joint Conference on Artificial Intelligence IJCAI 2007, Hyderabad, India, pp. 1513–1518 (January 2007)
7. Sierra, C., Debenham, J.: The LOGIC Negotiation Model. In: Proceedings Sixth International Conference on Autonomous Agents and Multi Agent Systems AAMAS 2007, Honolulu, Hawai'i, pp. 1026–1033 (May 2007)
8. Debenham, J., Sierra, C.: When Trust Is Not Enough. In: Huemer, C., Setzer, T., Aalst, W., Mylopoulos, J., Rosemann, M., Shaw, M.J., Szyperski, C. (eds.) EC-Web 2011. LNBIP, vol. 85, pp. 246–257. Springer, Heidelberg (2011)
9. Matt, P.A., Morge, M., Toni, F.: Combining statistics and arguments to compute trust. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, Richland, SC, pp. 209–216. International Foundation for Autonomous Agents and Multiagent Systems (2010)
10. Burnett, C., Norman, T.J., Sycara, K.: Bootstrapping trust evaluations through stereotypes. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, Richland, SC, pp. 241–248. International Foundation for Autonomous Agents and Multiagent Systems (2010)
11. Sierra, C., Debenham, J.: Information-based reputation. In: Paolucci, M. (ed.) First International Conference on Reputation: Theory and Technology (ICORE 2009), Gargonza, Italy, pp. 5–19 (2009)
12. MacKay, D.: Information Theory, Inference and Learning Algorithms. Cambridge University Press (2003)
13. Buechner, J., Tavani, H.: Trust and multi-agent systems: applying the "diffuse, default model" of trust to experiments involving artificial agents. Ethics and Information Technology 13(1), 39–51 (2011)