

# When Trust Is Not Enough

John Debenham<sup>1</sup> and Carles Sierra<sup>2</sup>

<sup>1</sup> QCIS, UTS, Broadway, NSW 2007, Australia  
debenham@it.uts.edu.au

<sup>2</sup> Institut d'Investigació en Intel·ligència Artificial - IIIA,  
Spanish Scientific Research Council, CSIC  
08193 Bellaterra, Catalonia, Spain  
sierra@iia.csic.es

**Abstract.** The degree of *trust* that an agent has for another is the strength of the agent's belief that the other will enact its commitments without variation. A strong sense of trust may be sufficient justification for one agent to sign a contract with another when all that matters is the possibility of variation between commitment and enactment. In non-trivial contracts the agents' information is typically asymmetric with each agent knowing more about its ability to vary its actions within its contractual constraints than the other. To enable an agent to deal with the asymmetry of information we propose two models. First, a *relationship model* that describes what one agent knows about another, *including* the belief that it has in the reliability of that information. Second an integrity model where *integrity* is the strength of an agent's belief that the other will not take advantage of its information asymmetries when enacting its commitments.

## 1 Introduction

The term *trust* is used in the sense of “certainty based on past experience”, and is commonly used particularly as the strength of belief that an agent has in another's desire to enact its commitments without variation. The literature on trust is enormous. 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. Trust is used here in line with the first approach where trust is something that is learned and evolves, although this does not mean that we view the second as less important [2]. *Reputation* is the opinion (more technically, a social evaluation) of a group about something — in a social environment. Reputation [3] feeds into trust. [4] presents a comprehensive categorisation of trust research: policy-based, reputation-based, and trust in information resources. [5] presents an interesting taxonomy of trust models in terms of nine types of trust models. [6] describes a powerful model that integrates interaction and role-based trust with witness and certified reputation.

Information asymmetry between contractually-bound agents has been studied extensively, and reached prominence with the award of the 2001 Nobel Prize in Economics to George Akerlof, Michael Spence, and Joseph E. Stiglitz “for their analyses of markets with asymmetric information.” Contract theory tackles information asymmetry by

invoking the unrealistic concept of a *complete contract* that specifies the consequences of every possible state of the world [7]. In real situations, agents accept that contracts are incomplete and rely on their contractual partner to ‘do the right thing’. In other words, an agent relies on them to act with integrity, where *integrity* is the strength of belief that the other will *not* take advantage of its information asymmetries when enacting his commitments. An agent will be (economically) motivated to act with integrity when it prefers to develop an on-going (business) relationship with another agent rather than taking full advantage of each opportunity as it occurs. An agent who exhibits this latter behaviour may need to continually seek new trading partners if past partners are not motivated to trade again. It is proposed that the development of a sense of integrity comes hand-in-hand with the development of (business) relationships. In particular, the estimation of integrity is predicated on the existence of a model of relationships.

This paper is concerned with tools to manage variations in agent behaviour that may take advantage of information asymmetries whilst being trustworthy, i.e. within its contractual commitments. Two tools are proposed. First, relationships described in Section 2, and an associated relationship model described in Section 3. Second, an integrity model described in Section 4. Section 5 concludes.

## 2 ‘Relationships’ between Agents

There is evidence from psychological studies that humans seek a *balance* in their negotiation relationships. The classical view [8] is that people perceive resource allocations as being distributively fair (i.e. well balanced) if they are proportional to inputs or contributions (i.e. equitable). However, more recent studies [9,10] show that humans follow a richer set of norms of distributive justice depending on their *intimacy* level: equity, equality, and need. *Equity* being the allocation proportional to the effort (e.g. the profit of a company goes to the stock holders proportional to their investment), *equality* being the allocation in equal amounts (e.g. two friends eat the same amount of a cake cooked by one of them), and *need* being the allocation proportional to the need for the resource (e.g. in case of food scarcity, a mother gives all food to her baby).

We believe that the perception of balance in dialogues (in negotiation or otherwise) is grounded on social relationships, and that every dimension of an interaction between humans can be correlated to the social closeness, or *intimacy*, between the parties involved. The more intimacy the more the need norm is used, and the less intimacy the more the equity norm is used. This might be part of our social evolution. There is ample evidence that when human societies evolved from a hunter-gatherer structure<sup>1</sup> to a shelter-based one<sup>2</sup> the probability of survival increased when food was scarce.

In this context, we can clearly see that, for instance, families exchange not only goods but also information and knowledge based on need, and that few families would consider their relationships as being unbalanced, and thus unfair, when there is a strong

<sup>1</sup> In its purest form, individuals in these societies collect food and consume it when and where it is found. This is a pure equity sharing of the resources, the gain is proportional to the effort.

<sup>2</sup> In these societies there are family units, around a shelter, that represent the basic food sharing structure. Usually, food is accumulated at the shelter for future use. Then the food intake depends more on the need of the members.

asymmetry in the exchanges (a mother explaining everything to her children, or buying toys, and then does not expect reciprocity). In the case of partners there is some evidence [11] that the allocations of goods and burdens (i.e. positive and negative utilities) are perceived as fair, or in balance, based on equity for burdens and equality for goods.

The perceived balance in a negotiation dialogue allows negotiators to infer information about their opponent, about its stance, and to compare their relationships with all negotiators. For instance, if we perceive that every time we request information it is provided, and that no significant questions are returned, or no complaints about not receiving information are given, then that probably means that our opponent perceives our social relationship to be very close. Alternatively, we can detect what issues are causing a burden to our opponent by observing an imbalance in their information or utilitarian utterances on that issue.

A *relationship* between two agents is somehow encapsulated in their *history* that is a complete record of their interactions. This potentially large amount of information is usually summarised by agents into various models. For example, the majority of agents construct a world model and a trust model. This paper is concerned with two models that are designed to assist an agent to deal with pervasive information asymmetry founded on the observation that each agent knows more about its own commitments and its intended enactments than any other agent. These two models are a relationship model described in Section 3 and an integrity model described in Section 4.

This Section describes the LOGIC illocutionary framework for classifying argumentative interactions. This framework was first described in [12] where it was used to help agents to prepare for a negotiation in the *prelude stage* of an interaction<sup>3</sup>. This paper generalises that framework and uses it to define one of the two dimensions of the relationship model described in Section 3, the second dimension is provided by the structure of the ontology<sup>4</sup>. The five LOGIC categories for information are quite general:

- Legitimacy contains *information* that may be part of, relevant to or in justification of contracts that have been signed.
- Options contains information about *contracts* that an agent may be prepared to sign.
- Goals contains information about the *objectives* of the agents.
- Independence contains information about the agent's *outside options* — i.e. the set of agents that are capable of satisfying each of the agent's needs.
- Commitments contains information about the *commitments* that an agent has.

and are used here to categorise all incoming communication that feeds into the agent's relationship model. As we will see this categorisation is not a one-to-one mapping and some illocutions fall into multiple categories. These categories are designed to provide a model of the agents' information that is relevant to their relationships, and are

<sup>3</sup> The five stages of an interaction dialogue are described in Section 4.

<sup>4</sup> All that we require of the ontology is that it has a partial order  $\leq$  defined by the is-a hierarchy, and a distance measure between concepts such as:  $\delta(c_1, c_2) = e^{-\kappa_1 l} \cdot \frac{e^{\kappa_2 h} - e^{-\kappa_2 h}}{e^{\kappa_2 h} + e^{-\kappa_2 h}}$  which is described in [13] where  $l$  is the shortest path between the concepts,  $h$  is the depth of the deepest concept subsuming both concepts, and  $\kappa_1$  and  $\kappa_2$  are parameters scaling the contribution of shortest path length and depth respectively.

not intended to be a universal categorising framework for all utterances. The LOGIC framework for managing communication is illustrated in Figure 1. A simplified formal model relates the LOGIC framework to the BDI model:

- $L = \{B(\alpha, \varphi)\}$ , that is a set of *beliefs*.
- $O = \{\text{Plan}(\langle \alpha_1, \text{Do}(p_1) \rangle, \dots, \langle \alpha_n, \text{Do}(p_n) \rangle)\}$ , that is a set of *joint plans*.
- $G = \{D(\alpha, \varphi)\}$ , that is a set of *desires*.
- $I = \{\text{Can}(\alpha, \text{Do}(p))\}$ , that is a set of *capabilities*.
- $C = \{I(\alpha, \text{Do}(p))\} \cup \{\text{Commit}(\alpha, \text{Do}(p))\}$ , that is a set of *commitments* and *intentions*.

This paper is written from the point of view of an agent  $\alpha$  is in a *multiagent system* with a finite number of other agents  $\mathcal{B} = \{\beta_1, \beta_2, \dots\}$ , and a finite number of *information providing agents*  $\Theta = \{\theta_1, \theta_2, \dots\}$  that provide the *context* for all events in the system —  $\Theta^t$  denotes the state of these agents at time  $t$ .  $\alpha$  observes the actions of another agent  $\beta$  in the context  $\Theta^t$ . The only thing that  $\alpha$  ‘knows for certain’ is its *history* of past communication that is retains in the repository  $\mathcal{H}_\alpha^t$ . Each *utterance* in the history contains: an illocutionary statement, the sending agent, the receiving agent, the time that the utterance was sent or received.

Observations are of little value unless they can be verified.  $\alpha$  may not possess a comprehensive range of reliable sensory input devices. Sensory inadequacy is dealt with invoking an *institution agent*,  $\xi$ , that truthfully, accurately and promptly reports what it sees. So if  $\beta$  commits to delivering twelve sardines at 6:00pm, or states that “it will rain tomorrow” and is committed to the truth of that prediction, then  $\alpha$  will eventually be in a position to verify those commitments when  $\xi$  advises what actually occurs.  $\xi$  is simply a convenient abstraction to deal with the problem of sensory inadequacy of software agents. As we will see below, agent  $\alpha$  qualifies all utterances received, including offers, information, arguments, with an epistemic probability representing  $\alpha$ ’s belief in their veracity.  $\xi$  is the only agent that  $\alpha$  believes is always truthful.

All communication is recorded in  $\alpha$ ’s history  $\mathcal{H}_\alpha^t$  that in time may contain a large amount of data. The majority of agent architectures include models that summarise the contents of  $\mathcal{H}^t$ ; for example, a *world model* and a *trust model*. In this paper we describe two models, a *relationship model* and an *integrity model* that are specifically designed to assist an agent to manage information asymmetries. To build the relationship model we will use the LOGIC framework to categorise the information in utterances received. That is,  $\alpha$  requires a categorising function  $v : U \rightarrow \mathcal{P}(\{\mathbf{L}, \mathbf{O}, \mathbf{G}, \mathbf{I}, \mathbf{C}\})$  where  $U$  is the set of utterances. The power set,  $\mathcal{P}(\{\mathbf{L}, \mathbf{O}, \mathbf{G}, \mathbf{I}, \mathbf{C}\})$ , is required as some utterances belong to multiple categories. For example, “I will not pay more for wine than the price that John charges” is categorised as both Option and Independence.

### 3 The Relationship Model $\mathcal{R}_{\alpha\beta}^t$

All of  $\alpha$ ’s models are summaries of its history  $\mathcal{H}_\alpha^t$ . The *relationship model* that  $\alpha$  has of  $\beta$  consists of four component models. First,  $\alpha$ ’s *intimacy model* of  $\beta$ ’s private information describes *how much*  $\alpha$  knows about  $\beta$ ,  $I_{\alpha\beta}^t$ . Second,  $\alpha$ ’s *reliability model* of *how*

*reliable* is the information summarised in  $I_{\alpha\beta}^t, R_{\alpha\beta}^t$ . Third,  $\alpha$ 's *reflection model* of  $\beta$ 's model of  $\alpha$ 's private information,  $J_{\alpha\beta}^t$ . Fourth, a *balance model*,  $B_{\alpha\beta}^t$ , that measures the difference in the rate of growth of  $I_{\alpha\beta}^t$  and  $J_{\alpha\beta}^t$ .

The remainder of this section details how these four models are calculated. This is achieved by extracting data from the process used to update the agent's world model  $\mathcal{M}^t$  — if an agent maintains the currency of their world model then the marginal cost of building these four models is low. The description given employs the machinery to update the world model in our information-based agents [14]. However it can be adapted to the machinery used by any agent that represents uncertainty in its world model using probability distributions, in which case  $\mathcal{M}^t = \{X_i\}_i$  where  $X_i$  are random variables. In addition to the world model and the models described in this paper an agent may construct other models such as a *trust model* and an *honour model* [15].

The idea of intimacy and balance is that intimacy summarises the degree of closeness, and *balance* is degree of fairness. Informally, *intimacy* measures how much one agent knows about another agent's private information, and *balance* measures the extent to which information revelation between the agents is 'fair'. The *intimacy* and *balance* models are structured using the LOGIC illocutionary framework and the ontology  $\mathcal{O}^5$ . For example, the communication  $\text{Accept}(\beta, \alpha, \delta)$  meaning that agent  $\beta$  accepts agent  $\alpha$ 's previously offered deal  $\delta$  is classified as an Option, and  $\text{Inform}(\beta, \alpha, \text{info})$  meaning that agent  $\beta$  informs  $\alpha$  about *info* and commits to the truth of it is classified as Legitimacy. The *balance model* of  $\alpha$ 's relationship with  $\beta$ ,  $B_{\alpha\beta}^t$ , is the element by element numeric difference of  $\frac{d}{dt}I_{\alpha\beta}^t$  and  $\frac{d}{dt}J_{\alpha\beta}^t$  across the structure  $\{\mathbf{L}, \mathbf{O}, \mathbf{G}, \mathbf{I}, \mathbf{C}\} \times \mathcal{O}$ .

### 3.1 The Components $I_{\alpha\beta}^t, R_{\alpha\beta}^t$ and $J_{\alpha\beta}^t$

The *intimacy* of  $\alpha$ 's relationship with  $\beta$ ,  $I_{\alpha\beta}^t$ , is the amount that  $\alpha$  knows about  $\beta$ 's private information and is represented as real numeric values over  $\{\mathbf{L}, \mathbf{O}, \mathbf{G}, \mathbf{I}, \mathbf{C}\} \times \mathcal{O}$ . Suppose  $\alpha$  receives utterance  $u$  from  $\beta$  and that the LOGIC category  $f \in v(u)$ , where  $v$  is the categorising function described above. For any concept  $c \in \mathcal{O}$ , define  $\Delta(u, c) = \max_{c' \in u} \delta(c', c)$  where  $\delta$  is a semantic distance function such as that described in Footnote 4. Denote the value of  $I_{\alpha\beta}^t$  in position  $(f, c) \in \{\mathbf{L}, \mathbf{O}, \mathbf{G}, \mathbf{I}, \mathbf{C}\} \times \mathcal{O}$  by  $I_{\alpha\beta(f,c)}^t$  then:

$$I_{\alpha\beta(f,c)}^t = \rho \times I_{\alpha\beta(f,c)}^{t-1} + (1 - \rho) \times \mathbb{I}^t(u) \times \Delta(u, c) \quad (1)$$

for any  $c$ , where  $\rho$  is the discount rate, and  $\mathbb{I}^t(u)$  is Shannon information gain as given by Equation 7.  $\alpha$ 's estimate of  $\beta$ 's intimacy on  $\alpha$ ,  $J_{\alpha\beta}^t$ , is constructed similarly by assuming that  $\beta$ 's reasoning apparatus mirrors  $\alpha$ 's.

The reliability model for utterance  $u$  is updated subsequent to the receipt of  $u$  when the institution agent  $\xi$  advises  $\alpha$  that  $u'$  was observed instead of  $u$  that was expected. Denote the value of  $R_{\alpha\beta}^t$  in position  $(f, c)$  by  $R_{\alpha\beta(f,c)}^t$  then:

$$R_{\alpha\beta(f,c)}^t = \rho \times R_{\alpha\beta(f,c)}^{t-1} + (1 - \rho) \times \mathbb{R}^t(u)|u' \times \Delta(u, c) \quad (2)$$

for any  $c$ , where  $\rho$  is the discount rate, and  $\mathbb{R}^t(u)|u'$  is given by Equation 9.

<sup>5</sup> Only a subset of the ontology is required. The idea is simply to capture "How much has Carles told me about wine", or "how much do it know about his commitments (possibly with other agents) concerning cheese".

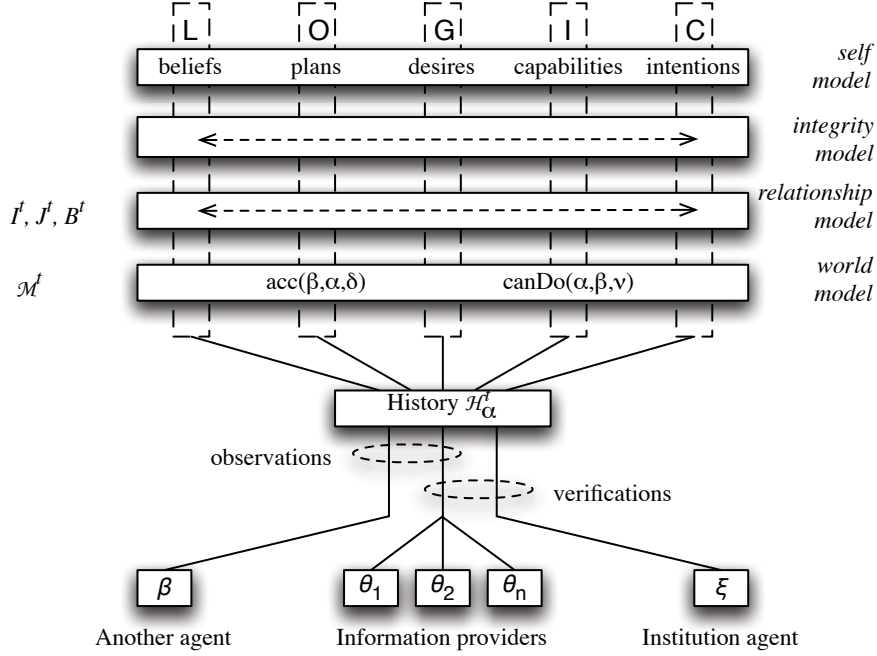


Fig. 1. The LOGIC framework for categorising information in an agent's relationship model

Utterances are represented in the world model  $\mathcal{M}_\alpha^t$  as probability distributions,  $(X_i)$ , in first-order probabilistic logic  $\mathcal{L}$ . For example, in a simple multi-issue contract negotiation  $\alpha$  may estimate  $\mathbb{P}^t(acc(\beta, \alpha, \delta))$ , the probability that  $\beta$  would accept contract  $\delta$ , by observing  $\beta$ 's responses. The distribution  $\mathbb{P}^t(acc(\beta, \alpha, \delta)) \in \mathcal{M}_\alpha^t$  is classified as an Option in LOGIC. Using shorthand notation, if  $\beta$  sends the message Offer( $\delta_1$ ) then  $\alpha$  may derive the constraint:  $K^{acc(\beta, \alpha, \delta)}(Offer(\delta_1)) = \{\mathbb{P}^t(acc(\beta, \alpha, \delta_1)) = 1\}$ , and if this is a counter offer to a former offer of  $\alpha$ 's,  $\delta_0$ , then:  $K^{acc(\beta, \alpha, \delta)}(Offer(\delta_1)) = \{\mathbb{P}^t(acc(\beta, \alpha, \delta_0)) = 0\}$ . In the not-atypical special case of multi-issue bargaining where the agents' preferences over the individual issues *only* are known and are complementary to each other's, maximum entropy reasoning can be applied to estimate the probability that any multi-issue  $\delta$  will be acceptable to  $\beta$  by enumerating the possible worlds that represent  $\beta$ 's "limit of acceptability" [14]. As another example, the predicate  $canDo(\alpha, \beta, \nu)$  meaning  $\beta$  is able to satisfy  $\alpha$ 's need  $\nu$  — this predicate is classified as Independence in LOGIC.

Updating  $\mathcal{M}_\alpha^t$  is complicated the need to take the integrity of utterances received into account — it would certainly be foolish for  $\alpha$  to believe that every utterance received from  $\beta$  was correct — whereas all utterances received from the institution agent  $\xi$  are assumed to be correct. The procedure for doing this, and for attaching an *a priori* belief to utterances (see Equation 10), is summarised in Section 3.3.

### 3.2 Estimating the information in an utterance: $\mathbb{I}^t(u)$

$\mathcal{M}_\alpha^t$  is a set of random variables,  $\mathcal{M}^t = \{X_i, \dots, X_n\}$  each representing an aspect of the world that  $\alpha$  is interested in. In the absence of in-coming messages the integrity of  $\mathcal{M}^t$  decays.  $\alpha$  may have background knowledge concerning the expected integrity as  $t \rightarrow \infty$ . Such background knowledge is represented as a *decay limit distribution*. One possibility is to assume that the decay limit distribution has maximum entropy whilst being consistent with observations. Given a distribution,  $\mathbb{P}(X_i)$ , and a decay limit distribution  $\mathbb{D}(X_i)$ ,  $\mathbb{P}(X_i)$  decays by:

$$\mathbb{P}^{t+1}(X_i) = \Delta_i(\mathbb{D}(X_i), \mathbb{P}^t(X_i)) \quad (3)$$

where  $\Delta_i$  is the *decay function* for the  $X_i$  satisfying the property that  $\lim_{t \rightarrow \infty} \mathbb{P}^t(X_i) = \mathbb{D}(X_i)$ . For example,  $\Delta_i$  could be linear:  $\mathbb{P}^{t+1}(X_i) = (1 - \nu_i) \times \mathbb{D}(X_i) + \nu_i \times \mathbb{P}^t(X_i)$ , where  $\nu_i < 1$  is the decay rate for the  $i$ 'th distribution. Either the decay function or the decay limit distribution could also be a function of time:  $\Delta_i^t$  and  $\mathbb{D}^t(X_i)$ .

The following procedure updates  $\mathcal{M}^t$  for all utterances  $u \in U$ . Suppose that  $\alpha$  receives a message  $u$  from agent  $\beta$  at time  $t$ . Suppose that this message states ‘‘If I were you then something is so’’ with probability  $z$ , and suppose that  $\alpha$  attaches an epistemic belief  $\mathbb{R}_{\alpha\beta}^t(u)$  to  $u$  — a method for estimating  $\mathbb{R}^t(u)$  is given below. Each of  $\alpha$ 's active plans,  $s$ , contains constructors for a set of distributions  $\{X_i\} \in \mathcal{M}^t$  together with associated *update functions*<sup>6</sup>,  $K_s(\cdot)$ , such that  $K_s^{X_i}(u)$  is a set of linear constraints on the posterior distribution for  $X_i$ . Denote the prior distribution  $\mathbb{P}^t(X_i)$  by  $\mathbf{p}$ , and let  $\mathbf{p}(u)$  be the distribution with minimum relative entropy<sup>7</sup> with respect to  $\mathbf{p}$ :  $\mathbf{p}(u) = \arg \min_{\mathbf{r}} \sum_j r_j \log \frac{r_j}{p_j}$  that satisfies the constraints  $K_s^{X_i}(u)$ . Then let  $\mathbf{q}(u)$  be the distribution:

$$\mathbf{q}(u) = \mathbb{R}_{\alpha\beta}^t(u) \times \mathbf{p}(u) + (1 - \mathbb{R}_{\alpha\beta}^t(u)) \times \mathbf{p} \quad (4)$$

and then let:

$$\mathbb{P}^t(X_{i(u)}) = \begin{cases} \mathbf{q}(u) & \text{if } \mathbf{q}(u) \text{ is ‘‘more interesting’’ than } \mathbf{p} \\ \mathbf{p} & \text{otherwise} \end{cases} \quad (5)$$

A general measure of whether  $\mathbf{q}(u)$  is *more interesting* than  $\mathbf{p}$  is:  $\mathbb{K}(\mathbf{q}(u) \parallel \mathbb{D}(X_i)) > \mathbb{K}(\mathbf{p} \parallel \mathbb{D}(X_i))$ , where  $\mathbb{K}(\mathbf{x} \parallel \mathbf{y}) = \sum_j x_j \ln \frac{x_j}{y_j}$  is the Kullback-Leibler distance between two probability distributions  $\mathbf{x}$  and  $\mathbf{y}$ .

Finally merging Equation 5 and Equation 3 we obtain the method for updating a distribution  $X_i$  on receipt of a message  $u$ :

$$\mathbb{P}^{t+1}(X_i) = \Delta_i(\mathbb{D}(X_i), \mathbb{P}^t(X_{i(u)})) \quad (6)$$

<sup>6</sup> A sample update function for the distribution  $\mathbb{P}^t(\text{acc}(\beta, \alpha, \delta))$  is given above.

<sup>7</sup> Given a probability distribution  $\mathbf{q}$ , the *minimum relative entropy distribution*  $\mathbf{p} = (p_1, \dots, p_I)$  subject to a set of  $n$  linear constraints  $\mathbf{g} = \{g_j(\mathbf{p}) = \mathbf{a}_j \cdot \mathbf{p} - c_j = 0\}$ ,  $j = 1, \dots, n$  (that must include the constraint  $\sum_i p_i - 1 = 0$ ) is:  $\mathbf{p} = \arg \min_{\mathbf{r}} \sum_j r_j \log \frac{r_j}{q_j}$ . This may be calculated by introducing Lagrange multipliers  $\lambda$ :  $L(\mathbf{p}, \lambda) = \sum_j p_j \log \frac{p_j}{q_j} + \lambda \cdot \mathbf{g}$ . Minimising  $L$ ,  $\{\frac{\partial L}{\partial \lambda_j} = g_j(\mathbf{p}) = 0\}$ ,  $j = 1, \dots, n$  is the set of given constraints  $\mathbf{g}$ , and a solution to  $\frac{\partial L}{\partial p_i} = 0$ ,  $i = 1, \dots, I$  leads eventually to  $\mathbf{p}$ . Entropy-based inference is a form of Bayesian inference that is convenient when the data is sparse [16] and encapsulates common-sense reasoning [17].

This procedure deals with integrity decay, and with two probabilities: first, the probability  $z$  in the utterance  $u$ , and second the belief  $\mathbb{R}_{\alpha\beta}^t(u)$  that  $\alpha$  attached to  $u$ . the Shannon information gain in  $X_i$  is:

$$\mathbb{I}^t X_i = \mathbb{H}^t(X_i) - \mathbb{H}^{t-1}(X_i)$$

and if the distributions in  $\mathcal{M}^t$  are independent then the Shannon information gain for  $\mathcal{M}^t$  following the receipt of utterance  $u$  is:

$$\mathbb{I}^t(u) = \sum_{X_i} \mathbb{I}^t X_i \quad (7)$$

### 3.3 Estimating the Reliability of an Utterance: $\mathbb{R}^t(u)$

$\mathbb{R}_{\alpha\beta}^t(u)$  is an epistemic probability that takes account of  $\alpha$ 's personal caution. An empirical estimate of  $\mathbb{R}_{\alpha\beta}^t(u)$  may be obtained by measuring the 'difference' between commitment and observation. Suppose that  $u$  is received from agent  $\beta$  at time  $t$  and is verified by the institution agent,  $\xi$ , as  $u'$  at some later time  $t'$ . Denote the prior  $\mathbb{P}^t(X_i)$  by  $\mathbf{p}$ . Let  $\mathbf{p}_{(u)}$  be the posterior minimum relative entropy distribution subject to the constraints  $K_s^{X_i}(u)$ , and let  $\mathbf{p}_{(u')}$  be that distribution subject to  $K_s^{X_i}(u')$ . We now estimate what  $\mathbb{R}_{\alpha\beta}^t(u)$  should have been in the light of knowing *now*, at time  $t'$ , that  $u$  should have been  $u'$ .

The idea of Equation 4, is that  $\mathbb{R}_{\alpha\beta}^t(u)$  should be such that, *on average* across  $\mathcal{M}^t$ ,  $\mathbf{q}_{(u)}$  will predict  $\mathbf{p}_{(u')}$  — no matter whether or not  $u$  was used to update the distribution for  $X_i$ , as determined by the condition in Equation 5 at time  $u$ . The *observed reliability* for  $u$  and distribution  $X_i$ ,  $\mathbb{R}X_i^t(u)|u'$ , on the basis of the verification of  $u$  with  $u'$ , is the value of  $k$  that minimises:

$$\mathbb{R}X_i^t(u)|u' = \arg \min_k \mathbb{K}(k \cdot \mathbf{p}_{(u)} + (1 - k) \cdot \mathbf{p} \parallel \mathbf{p}_{(u')})$$

where  $\mathbb{K}$  is the Kullback-Leibler distance. The predicted *information* in  $u$  with respect to  $X_i$  is:

$$\mathbb{I}X_i^t(u) = \mathbb{H}^t(X_i) - \mathbb{H}^t(X_{i(u)}) \quad (8)$$

that is the reduction in uncertainty in  $X_i$  where  $\mathbb{H}(\cdot)$  is Shannon entropy. Equation 8 takes account of the value of  $\mathbb{R}X_i^t(u)$ .

If  $\mathbf{X}(u)$  is the set of distributions in  $\mathcal{M}^t$  that  $u$  affects, then the *observed reliability* of  $\beta$  on the basis of the verification of  $u$  with  $u'$  is:

$$\mathbb{R}^t(u)|u' = \frac{1}{|\mathbf{X}(u)|} \sum_i \mathbb{R}X_i^t(u)|u' \quad (9)$$

For any concept  $c \in \mathcal{O}$ ,  $\mathbb{R}^t(c)$  is  $\alpha$ 's estimate of the reliability of information from  $\beta$  concerning  $c$ . In the absence of incoming communications the integrity of this estimate will decay in time by:  $\mathbb{R}^t(c) = \chi \times \mathbb{R}^{t-1}(c)$  for decay constant  $\chi < 1$  and close to 1. On receipt of communication  $u$  concerning  $c$  that is subsequently verified as  $u'$ :

$$\mathbb{R}^t(c) = \mu \times \mathbb{R}^{t-1}(c) + (1 - \mu) \times \mathbb{R}^t(u)|u' \quad (10)$$

where  $\mu$  is the learning rate, that estimates the reliability of  $\beta$ 's advice on any concept  $c$ . This completes the estimation of  $\mathbb{I}^t(u)$  and  $\mathbb{R}^t(u)$ .

#### 4 The Integrity Model $\mathcal{I}_{\alpha\beta}^t$

Agents interact through various forms of dialogues. This paper is concerned with *commitment dialogues* that contain at least one commitment, where a *commitment* may be to the truth of a statement or may be a contractual commitment. We assume that all commitment dialogues take place in some or all of the following five stages:

1. the *prelude* during which agents prepare for the interaction
2. the *negotiation* that may lead to
3. *signing* a contract
4. the *enactment* of the commitments in the contract
5. the *appraisal* of the complete interaction process that is made when the goods or services acquired by enactment of the contract have been consumed

The term *trust* is commonly used to refer to the enactment of commitments [4], and is evaluated at the completion of the enactment step in a commitment dialogue. ‘Integrity’ is distinct from trust, and is concerned with the appraisal of the dialogue including the behaviour of partner agents in commitment dialogues. For example, when ordering a bottle of wine, the merchant is *trustworthy* if the bottle is delivered as contractually specified, and the merchant will have acted with *integrity* if the wine is in good condition when it is consumed — possibly at a considerably later time.

The *integrity* of agent  $\beta$  is the strength of  $\alpha$ ’s belief that  $\beta$  will enact its contractual commitments so as to take account of  $\alpha$ ’s interests rather than executing the contract selfishly ‘to the letter’. For example, “I haven’t got the strawberries you ordered because yesterday’s hail storm is likely to have bruised the fruit”. Integrity is measured on a finite, fuzzy scale containing values such as ‘perfect’ and ‘terrible’. For some dialogues the appraisal stage may take place a considerable time after the enactment stage; for example, “Carles advised me to buy the Mercedes and I after three years I am still delighted with it” that implicitly rates the quality of Carles’ advice. This time delay is the reason that some business relationships necessarily take time to develop.

The integrity model is required to do the following. Given a particular need  $\nu$  and the prevailing contextual information  $\Theta^t$ ,  $\mathcal{I}_{\alpha\beta}^t$  aims to estimate the integrity of each agent in satisfying  $\nu$  on the basis of the past commitment dialogues recorded in  $\mathcal{H}_{\alpha}^t$ . From the set of commitment dialogues in  $\mathcal{H}_{\alpha}^t$  with agent  $\beta$ , we first form  $\mathcal{C}_{\alpha\beta}^t$  that contains: an abstraction of the need that triggered the dialogue, the prevailing contextual information and the resulting evaluation of the dialogue. The abstraction of the need  $\nu$  is to a chosen level using the  $\leq$  relation in the ontology — see Footnote 4. For example,  $\mathcal{H}_{\alpha}^t$  may contain a dialogue involving buying potatoes from  $\beta$  in which case  $\mathcal{C}_{\alpha\beta}^t$  could contain a record involving ‘root vegetables’ together with the contextual information that prevailed at that time, and the evaluation.

$\mathcal{I}_{\alpha\beta}^t$  aims to form beliefs on the evaluation of future commitment dialogues with agent  $\beta$  based on  $\mathcal{C}_{\alpha\beta}^t$  by treating the evaluations as values of the dependent variable. This can be interpreted as a pattern mining exercise from the information in  $\mathcal{C}_{\alpha\beta}^t$  to find the ‘best’ hypothesis that describes  $\mathcal{C}_{\alpha\beta}^t$ . One neat way to perform this induction is the

*minimum description length principle* [18] that is founded on the minimisation of the cost of communicating a body of knowledge from one agent to another that thus has a fundamental affinity with distributed autonomous systems:

$$\mathcal{I}_{\alpha\beta}^t \triangleq \arg \min_M (\mathbb{L}(M) + \mathbb{L}(\mathcal{C}_{\alpha\beta}^t | M)) \quad (11)$$

where  $\mathbb{L}(M)$  is the length of the shortest encoding of  $M$ , and  $\mathbb{L}(\mathcal{C}_{\alpha\beta}^t | M)$  is the length of the shortest encoding of  $\mathcal{C}_{\alpha\beta}^t$  given  $M$ . This definition is as neat as it is computationally expensive — it divides  $\mathcal{C}_{\alpha\beta}^t$  into that which may be generalised and that which may not.

The definition of  $\mathcal{I}_{\alpha\beta}^t$  in Equation 11 appears problematic for three reasons. First, if  $M$  can be any Turing computable model the definition is not computable, second there should be a language for representing  $M$ , and third the meaning of ‘the length of the shortest encoding’ is not clear. The second and third reason have been resolved [18]. The first, computability problem can be solved by restricting the models to some specific class. If the models are restricted to Bayesian decision graphs over finite spaces then Equation 11 is computable [19].

The model does not take time into account. In some applications old observations may be poorer indicators that recent ones, but this is not always so. To allow for varying strength of observations with time we construct instead  $\mathcal{C}_{\alpha\beta}^{*t}$  that is the same as  $\mathcal{C}_{\alpha\beta}^t$  except each appraisal,  $x$ , is replaced by a random variable  $X$  over appraisal space. These probability distributions are constructed by:  $\lambda \times X + (1 - \lambda) \times D_X$  where  $D_X$  is the *decay limit distribution*<sup>8</sup> for  $X$  — and  $X$  is a distribution with a ‘1’ indicating the position of the appraisal and 0’s elsewhere. This fine-grained approach gives control over the integrity decay of each observation.

Despite its elegance, Equation 11 is computationally expensive and we now describe methods for evaluating integrity for given  $\nu$  and  $\Theta^t$  for various  $\beta$ ’s. We represent the relationship between need  $\nu$ , context  $\Theta^t$  and appraisal  $a$  using conditional probabilities,  $\mathbb{P}_{\alpha\beta}^{t'}(a|\nu, \Theta^t)$ . If  $\nu$  is a need,  $\Theta^t$  the context that prevailed at the time  $t$  commitments were made, and  $a$  the resulting subsequent appraisal performed at time  $t'$  then  $\mathbb{P}_{\alpha\beta}^{t'}(a|\nu, \Theta^t)$  is the probability that  $a$  will be observed at time  $t'$  given that  $\beta$  had been selected to service need  $\nu$  in context  $\Theta^t$  at time  $t$ .

Any attempt to estimate  $\mathbb{P}_{\alpha\beta}^{t'}(a|\nu, \Theta^t)$  has to deal with the unbounded variation in context  $\Theta^t$ . We assume that there is a finite set of ‘essentially different’ contexts  $\Gamma$  and then estimate  $\mathbb{P}_{\alpha\beta}^t(a|\nu, \gamma)$  for  $\gamma \in \Gamma$ . Suppose that  $\mathbb{P}_{\alpha\beta}^t(a_i|\nu, \gamma)$  is observed where  $a_i \in A$  the finite appraisal space. Then  $\alpha$  attaches an epistemic strength  $d \in [0, 1]$  to this observation that is the probability that the same appraisal would be observed if the process was repeated for the same  $\nu$  and  $\gamma$ . Then  $\mathbb{P}_{\alpha\beta}^{t+1}(a|\nu, \gamma)$  is the distribution with minimum relative entropy to the prior  $\mathbb{P}_{\alpha\beta}^t(a|\nu, \gamma)$  subject to the constraint that  $\mathbb{P}_{\alpha\beta}^{t+1}(a_i|\nu, \gamma) = d$ .

In general it is desirable that observations should effect integrity estimates that are semantically close. This is achieved by appealing to a semantic similarity function,  $\delta$ , such as that described in Footnote 4, if observation  $\mathbb{P}_{\alpha\beta}^t(a_i|\nu', \gamma')$  is made with strength

<sup>8</sup> If the decay limit distribution is unknown we use a maximum entropy distribution.

$d'$  then the posterior for  $\mathbb{P}_{\alpha\beta}^{t+1}(a|\nu, \gamma)$  is the distribution with minimum relative entropy to the prior  $\mathbb{P}_{\alpha\beta}^t(a|\nu, \gamma)$  subject to the constraint that:

$$\mathbb{P}_{\alpha\beta}^{t+1}(a_i|\nu, \gamma) = \frac{b_i \times d''}{((1 - b_i) \times (1 - d'')) + (b_i \times d'')}, \text{ only if } d'' > 0.5$$

where  $d'' = d' \times \delta(\nu, \nu') \times \delta(\gamma, \gamma')$  discounts the effect of  $d'$  using  $\delta$ , and the condition  $d'' > 0.5$  limits the update region to  $\nu$  and  $\gamma$  that are semantically close to  $\nu'$  and  $\gamma'$  — this method assumes that the observations are independent. Then in the absence of new observations  $\mathbb{P}_{\alpha\beta}^t(a|\nu, \gamma)$  decays by Equation 3.

The estimate for  $\mathbb{P}_{\alpha\beta}^t(a|\nu, \gamma)$  in the previous paragraph enables  $\alpha$  to predict, or guess, the appraisal that will be observed if  $\alpha$  selects  $\beta$  to satisfy need  $\nu$  in context  $\gamma$ . It may be convenient to have a numeric score for  $\beta$ 's expected integrity given particular circumstances. One way to do this is to construct an 'ideal' distribution  $\mathbb{P}_I^t(a|\nu, \gamma)$  and to define integrity as the relative entropy between this ideal distribution and the estimated distribution:

$$G(\alpha, \beta, \nu, \gamma) = 1 - \sum_a \mathbb{P}_I^t(a|\nu, \gamma) \log \frac{\mathbb{P}_I^t(a|\nu, \gamma)}{\mathbb{P}_{\alpha\beta}^t(a|\nu, \gamma)}$$

A simpler way is to use a metricated, totally ordered appraisal space and to define integrity as expectation:  $G(\alpha, \beta, \nu, \gamma) = \sum_i a_i \times \mathbb{P}_{\alpha\beta}^t(a_i|\nu, \gamma)$ .

## 5 Discussion

This paper addresses the problem of dealing with information asymmetry that includes the observation that each agent knows more about its own commitments, and its intended enactments, than any other agent. Further agents may, and often do, deliberately conceal information to take strategic advantage. An agent can act in a perfectly trustworthy way, in the sense described above, whilst taking full advantage of the asymmetry of its information: "Well I did precisely what you asked me to do".

We have proposed two approaches to deal with information asymmetry. The first builds on the observation that in complex situations human agents prefer to interact with those with whom there is some depth of relationship to dealing with strangers. A relationship model has been described that measures the amount of private information that one agent knows about another, the reliability of that information and the balance in their information exchanges. Calculating these models is not simple, but substantially reuses those that update the agent's world model, and so the marginal cost of building the relationship model is small. The second approach models integrity that measures overall satisfaction with an interaction; it is updated at the appraisal stage that may be a considerable time after contract enactment.

Future work will focus on trialling the relationship model and the integrity model in a simulated marketplace. There can be no guarantee that an agent will act with integrity no matter how strong its relationships. So our goal will be to develop institutional incentives for agents to act with integrity based on published reputation measures, and then to show that the models described in this paper may be used to protect against unscrupulous exploitation of asymmetric information.

## References

1. Ramchurn, S., Huynh, T., Jennings, N.: Trust in multi-agent systems. *The Knowledge Engineering Review* 19, 1–25 (2004)
2. Arcos, J.L., Esteva, M., Noriega, P., Rodríguez, J.A., Sierra, C.: Environment engineering for multiagent systems. *Journal on Engineering Applications of Artificial Intelligence* 18 (2005)
3. Sabater, J., Sierra, C.: Review on computational trust and reputation models. *Artificial Intelligence Review* 24, 33–60 (2005)
4. 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, 58–71 (2007)
5. Viljanen, L.: Towards an ontology of trust. In: Katsikas, S., López, J., Pernum, G. (eds.) *Trust, Privacy and Security in Digital Business TrustBus 2005*, pp. 175–184. Springer, Heidelberg (2005)
6. Huynh, T., Jennings, N., Shadbolt, N.: An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 13, 119–154 (2006)
7. Bolton, P., Dewatripont, M.: *Contract Theory*. MIT Press, Cambridge (2005)
8. Adams, J.S.: Inequity in social exchange. In: Berkowitz, L. (ed.) *Advances in Experimental Social Psychology*, vol. 2. Academic Press, New York (1965)
9. Sondak, H., Neale, M.A., Pinkley, R.: The negotiated allocations of benefits and burdens: The impact of outcome valence, contribution, and relationship. *Organizational Behaviour and Human Decision Processes*, 249–260 (1995)
10. Valley, K.L., Neale, M.A., Mannix, E.A.: Friends, lovers, colleagues, strangers: The effects of relationships on the process and outcome of negotiations. In: Bies, R., Lewicki, R., Sheppard, B. (eds.) *Research in Negotiation in Organizations*, vol. 5, pp. 65–94. JAI Press (1995)
11. Bazerman, M.H., Loewenstein, G.F., White, S.B.: Reversal of preference in allocation decisions: judging an alternative versus choosing among alternatives. *Administration Science Quarterly*, 220–240 (1992)
12. 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 (2007)
13. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering* 15, 871–882 (2003)
14. Sierra, C., Debenham, J.: Information-based agency. In: *Proceedings of Twentieth International Joint Conference on Artificial Intelligence, IJCAI 2007*, Hyderabad, India, pp. 1513–1518 (2007)
15. Sierra, C., Debenham, J.: Trust and honour in information-based agency. In: Stone, P., Weiss, G. (eds.) *Proceedings Fifth International Conference on Autonomous Agents and Multi Agent Systems, AAMAS 2006*, pp. 1225–1232. ACM Press, Hakodate (2006)
16. Cheeseman, P., Stutz, J.: On The Relationship between Bayesian and Maximum Entropy Inference. In: *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, American Institute of Physics, Melville, NY, USA, pp. 445–461 (2004)
17. Paris, J.: Common sense and maximum entropy. *Synthese* 117, 75–93 (1999)
18. Grünwald, P.D.: *The Minimum Description Length Principle*. MIT Press, Cambridge (2007)
19. Suzuki, J.: Learning bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique. *IEICE Transactions on Information and Systems E81-D*, 356–367 (1998)