Trust and Honour in Information-based Agency

Carles Sierra
Institut d'Investigacio en Intel.ligencia Artificial
Spanish Scientific Research Council, UAB
08193 Bellaterra, Catalonia, Spain
sierra@iiia.csic.es

John Debenham
Faculty of Information Technology
University of Technology, Sydney
NSW, Australia
debenham@it.uts.edu.au

ABSTRACT

An argumentation based negotiation model is supported by information theory. Argumentative dialogues change the models of agents with respect to ongoing relationships between them. Trust and Honour are key components. Trust measures expected deviations of behaviour in the execution of commitments. Honour measures the expected integrity of the arguments exchanged. We understand the rhetorical moves in a dialogue as actions to project the current relationships into the future.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent systems*

General Terms

Theory

Keywords

Multiagent systems, Trust

1. INTRODUCTION

Negotiation dialogues have been traditionally organised around the basic illocutionary particles *Offer*, *Accept* and *Reject*. Previous work has been centred on the design of negotiation strategies and on proposing agent architectures able to deal with the exchange of offers [8, 4]. Game theory [13], possibilistic logic [5] and first-order logic [9] have been used for this purpose. Some initial steps in proposing rhetoric particles have been made, especially around the idea of *appeals*, *rewards* and *threats* [15]. However, no formal model of the meaning of these speech acts has been proposed yet. Expanded negotiation dialogues, including these rhetoric moves, are known as *argumentation-based negotiations*[1]. *Argumentation* in this sense is mainly to do with building (business) relationships. When we reward or threaten we refer to a future instant of time where the reward or threat will be effective, its scope goes beyond the current negotiation round. We will understand argumentation in this paper as an information exchange process between agents. Every illocution that an agent utters gives away valuable information. To evaluate each illocution exchanged we base our information-based agent architecture on information theory.

We distinguish between the *trust* that an agent displays through the enactment of its contracts, and the *honour* that an agent displays in sustaining its trading relationships. The enactment of a contract is uncertain to some extent, and trust, precisely, is a measure of how uncertain the enactment of a contract is. Trust is therefore a *measure of expected deviations of behaviour* along a dimension determined by the type of the contract [14]. In this sense, the higher the trust the lower the expectation that a (significant) deviation from what is signed occurs.

A (business) *relationship* between a set of agents consists of a historical sequence of related trades together with expectations of fair, trusted and honourable trade in the future. These expectations are established by the trading history and by the information exchanged [in appeals] and conditional promises made [in threats and rewards]. Exchanged information may turn out to be partly false, and promises may be partly broken. Honour measures the expected integrity of the information and promises exchanged. This leads to a natural hierarchy: individual deals, (business) relationships and businesses.

We first introduce an argumentation language in Section 2. Then, after giving our intuitions on building relationships is Section 3 we define an information based model in Section 4 that supports our model of Honour in Section 5. How the information gathered in argumentative dialogues is managed is explained in Section 6. Finally, the agent architecture is explained in Section 7.

2. AN ARGUMENTATION LANGUAGE

In order for communication to be effective agents need to agree upon an ontology and a communication language.

2.1 Ontology

To model the agent dialogues we define *ontology* to include a (minimum) repertoire of elements: a set of *concepts* organ-

[1]Not to be confused with argumentation in a classical sense: the generation of arguments, usually as logical proofs, for and against a given course of action that support decision making processes.

AAMAS'06, May 8-12, 2006, Hakodate, Japan.

ised in an *is-a hierarchy*, captured by a partial order relation, a set of *relations* defined over these concepts, and a set of axioms defined over the concepts and relations. We model ontologies following an algebraic approach and present ontologies as logical theories in the form of pairs $O = (S, A)$ where S is the *ontological signature*, representing the vocabulary, and A is a set of ontological axioms, specifying the interpreation of the vocabulary in a given context. The signature is usually represented as a mathematical structure and the set of axioms are usually written in a logical formalism. In our case, we define an ontology signature as a tuple $S = (C, R, \leq, \sigma)$ where C is a finite set of concept symbols (including basic data types); R is a finite set of relation symbols; $\leq$ is a reflexive, transitive and anti-symmetric relation on C (a partial order); and, $\sigma : R \to C^+$ is the function assigning to each relation symbol its arity. Concepts play somehow the role of *type*, and the is-a hierarchy is the notion of subtype. Thus, type inference mechanisms can be used to type all symbols appearing in expressions. Once the basic ontology signature is fixed we need to define a language to express, in our case, contracts, promises, rewards and so on. We denote that language as L_S. The axioms in the ontology will be written in that language, $A \subseteq L_S$, a first order language defined over the ontological expressions.

In our dialogues an essential concept is that of *contract*. Contracts are usually thought of as having a number of issues (dimensions) associated with values (or regions). Thus an (incomplete) example of ontology[2] for contracts could be defined as $\mathcal{S} = (C, R, \leq, \sigma)$ where

– $C = \{contract, quantity, integer, money, free, payment, date, plastic, wood, material, animal, crocodile\}$
– $R = \{price, made\_of, delivery, number, signed\}$
– $money \leq payment, free \leq payment, plastic \leq material, wood \leq material, crocodile \leq animal, integer \leq quantity$
– $\sigma(price) = contract \times payment, \sigma(made\_of) = contract \times material, \sigma(delivery) = contract \times date, \sigma(number) = contract \times quantity, \sigma(signed) = contract \times date$

For instance, the contract "3 plastic crocodiles at the price of €10" can be represented as:

$number(c_1, 3) \wedge made\_of(c_1, plastic) \wedge price(c_1, e10)$

A contract saying "3 crocodiles of any material, and at a maximum price of €10" becomes:

$number(c_1, 3) \wedge made\_of(c_1, material) \wedge price(c_1, [€0, €10])$

And, the expression "in all future contracts, I'll give you a free bottle of wine"

$\forall c : Contract\ .\ signed(c, [now, \infty]) \to present(c, bottle)$

We define the *ontological context* of formula $\varphi \in L_S$, denoted $O(\varphi)$, as the set of concepts in C used by the literals of φ. In the previous example, $O(\varphi) = \{contract, quantity, plastic, payment\}$. We extend the subtype relationship, $\leq$, in the following way,

$$\varphi \leq \psi \iff \forall c_i \in O(\varphi)\ (\exists c_j \in O(\psi).c_i \leq c_j)$$

Based on [10] we base the definition of the *semantic distance* between two concepts on the path length over the $\leq$ of the signature (more distance in the graph is more semantic distance), and the *depth* of the subsumer concept on the shortest path between the two concepts (the deeper in the hierarchy, the closer the meaning of the concepts). Semantic distance is then defined as:

[2]generic concepts and particular domains concepts are mixed in this example.

$$Sim(c_1, c_2) = e^{-\kappa_1 l} \cdot \frac{e^{\kappa_2 h} - e^{-\kappa_2 h}}{e^{\kappa_2 h} + e^{-\kappa_2 h}}$$

where l is the shortest path between the concepts, h is the depth of the deepest concept subsuming both concepts, and κ_1 and κ_2 are parameters scaling the contribution of shortest path length and depth respectively.

From the semantic distance between two concepts we define the semantic distance between formulae as the maxmin distance between concepts in their ontological context.

$$Sim(\varphi, \psi) = \max_{c_i \in O(\varphi)} \min_{c_j \in O(\psi)} \{Sim(c_i, c_j)\}$$

2.2 Communication language

Agent α is in a trading relationship with an agent β. They aim to strike deals $\delta = (a, b)$ where $a \in L_S$ is α's commitment and $b \in L_S$ is β's. Our agents share a communication language, $\mathcal{C}$, where the propositional content is expressed in L_S and the illocutionary particles are: [3]

– Offer(α, β, δ). Agent α offers agent β a deal $\delta = (a, b)$ with action commitments $a \in L_S$ for α and $b \in L_S$ for β.
– Accept(α, β, δ). Agent α accepts agent β's previously offered deal δ.
– Reject$(\alpha, \beta, \delta[, info])$. Agent α rejects agent β's previously offered deal δ. Optionally, information $info \in L_S$ explaining the reason for the rejection can be given.
– Withdraw$(\alpha, \beta[, info])$. Agent α breaks down negotiation with β. Extra $info \in L_S$ justifying the withdrawal may be given.
– Inform$(\alpha, \beta, info)$. Agent α informs β about $info \in L_S$ and commits to the truth of *info*.
– Reward$(\alpha, \beta, \delta, \varphi[, info])$. Intended to make the opponent accept a proposal with the promise of a future compensation. Agent α offers agent β a deal δ. In case β accepts the proposal, α commits to make $\varphi \in L_S$ true. The intended meaning is that α believes that worlds in which φ is true are somehow desired by β. Optionally, additional information in support of the deal can be given.
– Threat$(\alpha, \beta, \delta, \varphi, [info])$ Intended to make the opponent accept a proposal with the menace of some sort of retaliation. Agent α offers agent β a deal δ. In case β does not accept the proposal, α commits to make $\varphi \in L_S$ true. The intended meaning is that α believes that worlds in which φ is true are somehow not desired by β. Optionally, additional information in support of the deal can be given.
– Appeal$(\alpha, \beta, \delta, info)$ Intended to make the opponent accept a proposal as a consequence of the belief update that the accompanying information might bring about. Agent α offers agent β a deal δ. Additionally, α passes a pack of information in support of the deal. An Appeal can be understood as a combination of an offer and an inform, that is Appeal$(\alpha, \beta, \delta, info)$ = Offer(α, β, δ); Inform$(\alpha, \beta, info)$ — we borrow ';' from Dynamic Logic to mean action concatenation.

The accompanying information, *info*, can be of two basic types: (i) referring to the process (plan) used by an agent

[3]It is commonly accepted since the works by Austin and Searle that illocutionary acts are actions that succeed or fail. We will abuse notation in this paper and will consider that they are predicates in a first order logic meaning 'the action has been performed'. For those more pure-minded an alternative is to consider dynamic logic.

to solve a problem, or (ii) data (beliefs) of the agent including preferences. When building relationships, agents will therefore try to influence the opponent by changing their processes (plans) or by providing new data.

Dialogues, especially in Electronic Institutions [3], tend to be structured in order to facilitate the decision making of the participants. This structuring of dialogues is what is usually called a *protocol*. We will not detail it here, but will assume that such a protocol exists.

2.3 Information model predicates

The predicate $\text{Acc}(\alpha, \beta, \delta)$ means "agent α finds the deal δ with agent β acceptable", and $\text{Build}(\alpha, \beta, \rho)$ means "agent α considers agent β to be a potential trading partner for deals in a relationship ρ". Where ρ is a super type (more abstract set of concepts) than δ, that is $\delta \leq \rho$. ie: a relationship ρ could be "supply of chops", and a particular deal δ could be "supply 6 lamb chops today". Our agent estimates probabilities that are attached to $P(\text{Build}(\alpha, \beta, \rho))$ representing the certainty that it has in this proposition. Thus, if $P(\text{Build}(\alpha, \beta, \rho)) > P(\text{Build}(\alpha, \gamma, \rho))$ then it does *not* mean that β is better than γ. It *does* mean that for ρ-deals α is more certain that β is OK than that γ is OK.

To estimate $P(\text{Build}(\cdot))$ we express $\text{Build}(\cdot)$ in terms of a *social model* comprising:

– $\text{Capable}(\alpha, \beta, \rho)$ is a measure of the ability of an agent to do what it says it can do for deals of type ρ. Eg: "If β *says* he can supply both sausages and chops then he *can* supply both sausages and chops.".

– $\text{Trust}(\alpha, \beta, \rho)$ is a measure of expected deviations of behaviour along a given dimension (ρ). The assessment of Trust incorporates $\text{Reputation}(\beta)$ = "the reputation of β reported by other agents — not necessarily truthfully".

– $\text{Honour}(\alpha, \beta, \rho)$ measures the expected integrity of the arguments exchanged. That is, α's expectation that "β will honour previously issued appeals, threats and rewards when negotiating deals of type ρ". We expect rewards to compensate for bad proposals, and threats when our own position is seen to be dishonourable.

– $\text{Rel}(\alpha, \beta, c)$ measures the reliability of the information that β provides — this measure is keyed to sections of the ontological context, $c \in C$. For example, how reliable is β's information about wine, or about the stock market.

3. BUILDING RELATIONSHIPS

A simple bargaining agent α can only do three things: submit and respond to proposals, and "walk away". The only way in which a simple bargaining agent α can "reward or threaten" is to submit proposals that are more or less attractive to β than α would normally do so. A negotiation can "breakdown", and α may be interested in estimating $p_{b,\beta}$ — the probability that β will walk away in the subsequent negotiation round.

An argumentation agent extends the capacity of a simple bargaining agent by using threats, rewards and appeals to sustain trading relationships. An argumentation agent is concerned both with future single trades and with *trading relationships* that encapsulate expectations of future acceptable trades. Just as a negotiation can "breakdown", a trading relationship can *collapse*. So, just as a bargaining agent is interested in $p_{b,\beta}$, an argumentation agent will also be interested in estimating the probability that a trading relationship with β will collapse: $p_{c,\beta}$.

Our agent α uses argumentation to maintain a reliable set of potential trading partners, $\{\beta_i\}_{i \in \mathcal{B}_\rho}$, for each type of deal ρ. α considers β to be a reliable trading partner for ρ-deals if $P(\text{Build}(\alpha, \beta, \rho)) \geq \epsilon$ for some threshold constant ϵ. $\mathcal{B}_\rho = \{\beta \mid P(\text{Build}(\alpha, \beta, \rho)) \geq \epsilon\}$ will have an optimal size which is difficult to specify analytically due to dependencies between the probabilities. The size of the set $\mathcal{B}_\rho$ will be bounded by: first the difficulty in locating β_i for whom the p_{c,β_i} are "moderately" independent, and second the larger $\mathcal{B}_\rho$ the less opportunity α has to estimate $\{p_{c,\beta_i}\}_{i \in \mathcal{B}_\rho}$ directly. α may maintain these estimates indirectly by sharing information in a *trading pact*, but if α shares information about her trading partners then that may reduce the value of those partners to α.

Under what circumstances will α decide to "walk away" from a negotiation or a trading relationship? This is dealt with in α's negotiation plans, and α's plans for choosing potential trading partners — these are not described here. Presumably α will exercise her outside options if she believes that they will improve her position. α will only do this if $P(\text{Build}(\cdot))$ for her current partners are worse than the expected value for an untried partner. An argumentation agent's threat, reward and appeal illocutions are its weapons against collapse, as we discuss in Section 7.1.

4. AN INFORMATION-BASED MODEL

We ground our argumentation model on information-based concepts. *Entropy*, H, is a measure of uncertainty [11] in a probability distribution for a discrete random variable X: $H(X) \triangleq -\sum_i p(x_i) \log p(x_i)$ where $p(x_i) = P(X = x_i)$. Maximum entropy inference and minimum relative entropy inference are chosen partly because of their encapsulation of common sense reasoning [12].

Maximum entropy inference is used to derive sentence probabilities for that which is not known by constructing the "maximally noncommittal" [7] probability distribution, and minimum relative entropy inference is used to update these distributions. These forms of inference are criticised [6] for their dependence on the representation chosen — such as the way in which values for a continuous variable are represented as intervals. We argue to the contrary, that this choice enables the tailoring of the model in fine detail.

Given a prior probability distribution $\underline{q} = (q_i)_{i=1}^n$ and a set of constraints, the *principle of minimum relative entropy* chooses the posterior probability distribution $\underline{p} = (p_i)_{i=1}^n$ that has the least relative entropy with respect to $\underline{q}$, and that satisfies the constraints[4].

[1] describes the estimation of both $P(Acc(\alpha, \beta, \delta))$ and the estimation of $P(Acc(\beta, \alpha, \delta))$ which is α's estimate of β's willingness to accept δ. These estimates are derived by applying maximum entropy inference to the observed behaviour of the agents. In the subsequent subsection we'll see how α updates its sentence probabilities for $\text{Honour}(\cdot)$ by observation following receipt of the illocutionary particles described in Section 2.

[4]Ie: $\arg\min_{\underline{p}} \sum_{i=1}^n p_i \log \frac{p_i}{q_i}$. The principle of minimum relative entropy is a generalisation of the principle of maximum entropy. If the prior distribution $\underline{q}$ is uniform, the relative entropy of $\underline{p}$ with respect to $\underline{q}$ differs from $-H(\underline{p})$ only by a constant. So the principle of maximum entropy is equivalent to the principle of minimum relative entropy with a uniform prior distribution.

4.1 Updating honour from observations

The choice of trading partner is influenced by the observation of the fulfilment of informs, rewards and threats. Such illocutions are a *conditional commitment* to act. Eg: Reward$(\alpha, \beta, \delta, \varphi)$ when uttered by α has the meaning that α becomes socially committed to make φ happen if β enacts the illocution Accept(β, α, δ). Similarly, Threat$(\alpha, \beta, \delta, \varphi)$ when uttered by α has the meaning that α becomes socially committed to make φ happen if β does *not* enact the illocution Accept(β, α, δ) — usually by a given deadline. Finally, Inform(α, β, φ) means that α is socially committed to the truth of φ.

Agent α has the opportunity to observe the extent to which agent β sticks to his word in Reward$(\cdot)$, Threat$(\cdot)$ and Inform$(\cdot)$ illocutions, and to observe how accurate his *info* is. We base our measure of honour as the negative entropy of the probability distribution of possible outcomes following such a given illocution — honour measures the relationship between commitment and execution of promises. More precisely, between issued illocutions and their *perceived* execution. In this way, a natural way to base our modelling of honour is on a conditional probability, P^t, between commitment and observation given a context e as:

$$P^t(\text{Observe}(\alpha, \varphi') \mid (\text{Reward}(\beta, \alpha, \delta, \varphi); \text{Accept}(\alpha, \beta, \delta)), e)$$

where every reward execution represents a point in that distribution. For threats and informs the distributions are:

$$P^t(\text{Observe}(\alpha, \varphi') \mid (\text{Threat}(\beta, \alpha, \delta, \varphi); \neg\text{Accept}(\alpha, \beta, \delta)), e)$$

$$P^t(\text{Observe}(\alpha, \varphi') \mid (\text{Inform}(\beta, \alpha, \varphi)), e)$$

For simplicity, we denote these relationships between the commitment, φ, and the observation, φ', as $P^t(\varphi'|\varphi, e)$.

An important aspect that we want to model is the fact that beliefs 'evaporate' as time goes by. If we don't keep an ongoing relationship, we become unsure how *honourable* a trading partner is. If I stop buying meat from my butcher, I'm not sure anymore that he will commit to his promises. This decay is what justifies a continuous relationship between individuals. In our model, the conditional probabilities should tend to ignorance as represented by the *decay limit distribution* $\{d_i\}$. If we have the set of observations $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_n\}$ then complete ignorance of the opponent's expected behaviour means that given the opponent commits to φ the conditional probability for each observable outcome φ' becomes $d_i = \frac{1}{n}$ — i.e. the unconstrained maximum entropy distribution — but α may be prepared to make assumptions about β's decay limit distribution so that its entropy is less than this. This natural decay of belief is offset by new observations.

We define the evolution of the probability distribution that supports the previous definition of decay using an equation inspired by pheromone like models [2]:

$$P^{t+1}(\phi'|\phi) = \kappa \cdot \left((1-\nu) \cdot d_i + \nu \cdot \left(P^t(\phi'|\phi) + \Delta^t P(\phi'|\phi)\right)\right) \tag{1}$$

where κ is a normalisation constant and $\{d_i\}$ is the decay limit distribution for β. This equation models the passage of time for a conveniently large $\nu \in [0, 1]$ and where the term $\Delta^t P(\phi'|\phi)$ represents the increment in an instant of time according to the last experienced event as the following shows.

Similarity based. The question is how to use the observation φ' given a commitment φ (generated by a sequence of illocutions Reward$(\alpha, \beta, \delta, \varphi)$; Accept$(\beta, \alpha, \delta)$ — similarly for others) in the context of a deal $\delta \leq \rho$ in the update of the overall probability distribution over the set of all possible outcomes defined over L_C. Here we use the idea that given a particular deviation in a region of the space, *similar* deviations should be expected in other regions. The intuition behind the update is that if my butcher has not honoured commitments with respect to red meat, then I might expect similar deviations with respect to poultry.

This idea is built upon the previously defined *Sim* function (see Section 2) to take into account the difference between acceptance probabilities and similarity between the perception of the execution x of a reward (or threat) y. Thus, after the observation of φ' the increment of probability distribution at time $t+1$ is:

$$\Delta^t P(\phi'|\phi) = (1- \mid Sim(\varphi', \varphi) - Sim(\phi', \phi) \mid) \tag{2}$$

Entropy based. Suppose that α has a business relationship ρ with agent β, that β promises ϕ, and this promise is sound. The material value of ϕ to ρ will depend on the future use that α makes of it, and that is unlikely to be known. So α estimates the value of ϕ to the honour he holds for β in ρ using a probability distribution $(p_1, \ldots, p_n)$ over a relationship evaluation space $E = (e_1, \ldots, e_n)$ that could range from "that is what I expect from the perfect trading partner" to "it is totally useless" — E may contain hard or fuzzy values. $p_i = w_i(\rho, \phi)$ is the probability that e_i is the correct evaluation of the enactment ϕ in the context of relationship ρ, and $\underline{w} : L_S \times L_S \to [0, 1]^n$ is the promise *evaluation function*.

Let $L_S = (\phi_1, \ldots, \phi_m)$ in some order. Then for a given ϕ_k, $(P^t_\beta(\phi_1|\phi_k), \ldots, P^t_\beta(\phi_m|\phi_k))$ is the prior distribution of α's estimate of what will actually occur if β promised that ϕ_k would occur and $\underline{w}(\rho, \phi_k) = (w_1(\rho, \phi_k), \ldots, w_n(\rho, \phi_k))$ is α's evaluation over E with respect to the relationship ρ of β's promise ϕ_k. α's expected evaluation of what will occur given that β has promised that ϕ_k will occur is: $\underline{w}^{\exp}(\rho, \phi_k) =$

$$\left(\sum_{j=1}^{m} P^t_\beta(\phi_j|\phi_k) \cdot w_1(\rho, \phi_j), \ldots, \sum_{j=1}^{m} P^t_\beta(\phi_j|\phi_k) \cdot w_n(\rho, \phi_j)\right).$$

Now suppose that α observes the event $(\varphi'|\varphi)$ in another relationship ρ' also with agent β. Eg: α may buy wine and cheese from the same supplier. α may wish to revise the prior estimate $\underline{w}^{\exp}(\rho, \phi_k)$ in the light of the observation $(\varphi'|\varphi)$ to: $(\underline{w}^{\text{rev}}(\rho, \phi_k) \mid (\varphi'|\varphi)) =$

$$\underline{g}(\underline{w}^{\exp}(\rho, \phi_k), \underline{w}(\rho', \varphi), \underline{w}(\rho', \varphi'), \rho, \rho', \phi, \varphi, \varphi'),$$

for some function $\underline{g}$ — the idea being, for example, that if the promise, φ, concerning the purchase of cheese, ρ', was not kept then α's expectation that the promise, ϕ, concerning the purchase of wine, ρ, will not be kept should increase. The entropy based approach estimates $\Delta^t_\beta P(\phi'|\phi)$ by applying the principle of minimum relative entropy[4]. Let:

$$\left(P^t_{\beta,C}(\phi_j|\phi)\right)_{j=1}^m = \arg\min_{\underline{p}} \sum_{i=1}^{m} p_i \log \frac{p_i}{P^t_\beta(\phi_i|\phi)} \tag{3}$$

satisfying the n constraints C, and $\underline{p} = (p_j)_{j=1}^m$. Then:

$$\Delta^t P(\phi'|\phi) = P^t_{\beta,C}(\phi'|\phi) - P^t_\beta(\phi'|\phi) \tag{4}$$

Where the n constraints C are: $\sum_{j=1}^m p_j \cdot w_i(\rho, \phi_j) =$

$$g_i(\underline{w}^{\exp}(\rho, \phi_k), \underline{w}(\rho', \varphi), \underline{w}(\rho', \varphi'), \rho, \rho', \phi, \varphi, \varphi')$$

for $i = 1, \ldots, n$. This is a set of n linear equations in m unknowns, and so the calculation of the minimum relative entropy distribution may be impossible if $n > m$. In this case, we take only the m equations for which the change from the prior to the posterior value is greatest. That is, we attempt to select the most significant factors.

5. AN HONOUR MODEL

Honour as expected behaviour. Consider a distribution of expected fulfilment of promises that represent α's "ideal" for a relationship with β, in the sense that it is the best that α could reasonably expect β to do. This distribution will be a function of β, α's history with β, anything else that α believes about β, and general environmental information including time — denote all of this by e, then we have $P^t_I(\varphi'|\varphi, e)$. For example, if α considers that it is unacceptable for the execution φ' to be less preferred than the promise φ then $P^t_I(\varphi'|\varphi, e)$ will only be non-zero for those φ' that α prefers to φ. The distribution $P^t_I(\cdot)$ represents what α expects, or hopes, β will do. Honour is the relative entropy between this ideal distribution, $P^t_I(\varphi'|\varphi, e)$, and the distribution of the observation of fulfilled promises, $P^t_\beta(\varphi'|\varphi)$. That is:

$$H(\alpha, \beta, \varphi) = 1 - \sum_{\varphi': Sim(\varphi', \varphi) \geq \eta} P^t_I(\varphi'|\varphi, e) \log \frac{P^t_I(\varphi'|\varphi, e)}{P^t_\beta(\varphi'|\varphi)}$$

η being a minimum degree of similarity. This equation defines honour for one, single promise φ — for example, my honour in my butcher if he promises me a 10% discount for the rest of the year. It makes sense to aggregate these values over a class of promises, say over those φ that are subtypes of a particular relationship ρ, that is $\varphi \leq \rho$. In this way we measure the honour that I have in my butcher in relation to the promises he makes for red meat generally:

$$H(\alpha, \beta, \rho) = 1 - \frac{\sum_{\varphi:\varphi \leq \rho} P^t_\beta(\varphi) \left[\sum_{\varphi': Sim(\varphi', \varphi) \geq \eta} P^t_I(\varphi'|\varphi, e) \log \frac{P^t_I(\varphi'|\varphi, e)}{P^t_\beta(\varphi'|\varphi)}\right]}{\sum_{\varphi:\varphi \leq \rho} P^t_\beta(\varphi)}$$

where $P^t_\beta(\varphi)$ is a probability distribution over the space of promises that the next promise β will make to α is φ. Similarly, for an overall estimate of α's honour in β:

$$H(\alpha, \beta) = 1 - \sum_\varphi P^t_\beta(\varphi) \left[\sum_{\varphi': Sim(\varphi', \varphi) \geq \eta} P^t_I(\varphi'|\varphi, e) \log \frac{P^t_I(\varphi'|\varphi, e)}{P^t_\beta(\varphi'|\varphi)}\right]$$

Honour as expected 'selectability'. The previous notion of honour as expected behaviour was expressed in terms of our expected behaviour in an opponent that was defined for each promise φ. That notion requires that an ideal distribution, $P^t_I(\varphi'|\varphi, e)$, has to be specified for each φ. The specification of ideal distributions may be avoided by considering "expected selectability" of β instead of β's "expected behaviour". The idea is that in a trading relationship ρ we honour β more if the execution of a promise φ leads us to select him as a trading partner with greater confidence than if his promise had been executed precisely. Defining $P^t(\text{ObBuild}(\alpha, \beta, \rho, \varphi', \varphi)) = P^t(\text{Build}(\alpha, \beta, \rho)) \mid \text{Observe}(\varphi' \mid \varphi)$, then define:

$$f(x) = \begin{cases} 0 & \text{if } x < P^t(\text{ObBuild}(\alpha, \beta, \rho, \varphi, \varphi)) \\ 1 & \text{if } P^t(\text{ObBuild}(\alpha, \beta, \rho, \varphi, \varphi)) \\ & < x < P^t(\text{ObBuild}(\alpha, \beta, \rho, \varphi, \varphi)) + \epsilon \\ 0 & \text{otherwise} \end{cases}$$

(or perhaps a similar function with smoother shape) for some small ϵ, then define:

$$H(\alpha, \beta, \rho) = \frac{\sum_{\varphi:\varphi \leq \rho} P^t_\beta(\varphi) \cdot h(\alpha, \beta, \rho, \varphi)}{\sum_{\varphi:\varphi \leq \rho} P^t_\beta(\varphi)}$$

$$H(\alpha, \beta) = \sum_\varphi P^t_\beta(\varphi) \cdot h(\alpha, \beta, \rho, \varphi)$$

where $h(\alpha, \beta, \rho, \varphi) =$

$$\sum_{\varphi': Sim(\varphi', \varphi) \geq \eta} f(P^t(\text{ObBuild}(\alpha, \beta, \rho, \varphi', \varphi))) \cdot P^t_\beta(\varphi'|\varphi)$$

Honour as certainty in promise execution. Honour is consistency in expected acceptable executions of promises, or "the lack of expected uncertainty in those possible executions that are better than the promise as specified". The idea here is that α will honour β more if variations, φ', from expectation, φ, are not random. The Honour that an agent α has on agent β with respect to the execution of a promise φ is:

$$H(\alpha, \beta, \varphi) = 1 + \frac{1}{B^*} \cdot \sum_{\varphi': Sim(\varphi', \varphi) \geq \eta} P^t_+(\varphi'|\varphi) \log P^t_+(\varphi'|\varphi)$$

where $P^t_+(\varphi'|\varphi)$ is the normalisation of $P^t_\beta(\varphi'|\varphi)$ for those values of φ' for which

$$P^t(\text{ObBuild}(\alpha, \beta, \rho, \varphi', \varphi)) > P^t(\text{ObBuild}(\alpha, \beta, \rho, \varphi, \varphi))$$

and zero otherwise, $B(\varphi)^+$ is the nonempty set of executions of promises that α prefers to φ,

$$B^* = \begin{cases} 1 & \text{if } |B(\varphi)^+| = 1 \\ \log |B(\varphi)^+| & \text{otherwise} \end{cases}$$

As above we aggregate this measure for those promises of a particular type ρ:

$$H(\alpha, \beta, \rho) = 1 + \frac{\sum_{\varphi:\varphi \leq \rho} \left[P^t_\beta(\varphi) \cdot \sum_{\varphi': Sim(\varphi', \varphi) \geq \eta} P^t_+(\varphi'|\varphi) \log P^t_+(\varphi'|\varphi)\right]}{B^* \cdot \sum_{\varphi:\varphi \leq \rho} P^t_\beta(\varphi)}$$

$$= 1 + \frac{\sum_{\varphi:\varphi \leq \rho} \sum_{\varphi': Sim(\varphi', \varphi) \geq \eta} \left[P^t_+(\varphi', \varphi) \log P^t_+(\varphi'|\varphi)\right]}{B^* \cdot \sum_{\varphi:\varphi \leq \rho} P^t_\beta(\varphi)}$$

where $P^t_\beta(\varphi', \varphi)$ is the joint probability distribution, and $P^t_+(\varphi', \varphi)$ is the normalisation of it as above. And, similarly as before:

$$H(\alpha,\beta) =$$

$$1 + \frac{\sum_{\varphi}\left[P^t_{\beta}(\varphi)\cdot\sum_{\varphi':Sim(\varphi',\varphi)\geq\eta}P^t_{+}(\varphi'|\varphi)\log P^t_{+}(\varphi'|\varphi)\right]}{B^*}$$

$$= 1 + \frac{\sum_{\varphi}\sum_{\varphi':Sim(\varphi',\varphi)\geq\eta}\left[P^t_{+}(\varphi',\varphi)\log P^t_{+}(\varphi'|\varphi)\right]}{B^*}$$

6. DEALING WITH [*INFO*].

In this section we discuss how to deal with the information that β communicates in the illocutions, $\mathit{info} \in L_C$. We will base this treatment on β's *reliability* as an estimate of the extent to which this *info* is correct. For example, β may send α the *info* that "the price of fish will go up by 10% next week", and it may actually go up by 9%. α's argumentation and relationship building strategies are based on plans that are not described here. Those plans determine a set of probability distributions from which a world model is derived. *info* is represented as a set of linear constraints on one or more of those probability distributions. A chunk of *info* may not be directly related to one of those distributions, or may not be expressed naturally as constraints, and so some inference machinery is required to derive these constraints — this inference is performed by a set of model building function, J_s, for each active plan s chosen by α. If a plan s calls for the distribution D then J^D_s is the model building function for D, and $J^D_s(\mathit{info})$ is the set of constraints on D derived from *info*.

6.1 Updating the world model with *info*

If at time u, α receives a message containing *info* it is time-stamped and source-stamped $\mathit{info}_{(\beta,\alpha,u)}$, and placed in a repository $\mathcal{Y}^t$. If α has an active plan, s, then the model building function, J_s, is applied to $\mathit{info}_{(\beta,\alpha,u)}$ to derive constraints on some, or none, of α's distributions. The extent to which those constraints are permitted to effect the distributions is determined by a value for the *reliability* of β, $R^t(\alpha,\beta,O(\mathit{info}))$.

An agent may have models of integrity decay for some particular distributions, but general models of integrity decay for, say, a chunk of information taken at random from the World Wide Web are generally unknown. However the values to which decaying integrity should tend in time *are* often known. For example, a prior value for the truth of the proposition that a "22 year-old male will default on credit card repayment" is well known to banks. As described in Sec. 4.1, if α attaches such prior values to a distribution D they are called the *decay limit distribution* for D, $(d^D_i)^n_{i=1}$. No matter how integrity of *info* decays, in the absence of any other relevant information it should decay to the decay limit distribution.

In the absence of new *info* the integrity of distributions decays. If $D = (q_i)^n_{i=1}$ then we use a geometric model of decay:

$$q^{t+1}_i = (1-\nu^D)\times d^D_i + \nu^D \times q^t_i, \text{ for } i = 1,\ldots,n \tag{5}$$

where $\nu^D \in (0,1)$ is the decay rate. This raises the question of how to determine ν^D. Just as an agent may know the decay limit distribution it may also know something about ν^D. In the case of an information-overfed agent there is no harm in conservatively setting ν^D "a bit on the low side" as the continually arriving *info* will sustain the estimate for D.

We now describe how new *info* is imported to the distributions. A single chunk of *info* may affect a number of distributions. Suppose that a chunk of *info* is received from β and that α attaches the epistemic belief probability $R^t(\alpha,\beta,O(\mathit{info}))$ to it. Each distribution models a facet of the world. Given a distribution $D = (q_i)^n_{i=1}$, q_i is the probability that the possible world ω_i for D is the true world for D. The effect that a chunk *info* has on distribution D is to enforce the set of linear constraints on D, $J^D_s(\mathit{info})$. If the constraints $J^D_s(\mathit{info})$ are taken by α as valid then α could update D to the posterior distribution $(p^{\mathit{info}}_i)^n_{i=1}$ that is the distribution with least relative entropy with respect to $(q^t_i)^n_{i=1}$ satisfying the constraint:

$$\sum_i \{p^{\mathit{info}}_i : J^D_s(\mathit{info}) \text{ are all } \top \text{ in } \omega_i\} = 1. \tag{6}$$

But $R^t(\alpha,\beta,O(\mathit{info})) = r \in [0,1]$ and α should only treat the $J^D_s(\mathit{info})$ as valid if $r = 1$. In general r determines the extent to which the effect of *info* on D makes it closer to $(p^{\mathit{info}}_i)^n_{i=1}$ or to the prior $(q^t_i)^n_{i=1}$ distribution by:

$$p^t_i = r \times p^{\mathit{info}}_i + (1-r)\times q^t_i \tag{7}$$

But, we should only permit a new chunk of *info* to influence D if doing so gives us new information. For example, if 5 minutes ago a trusted agent advises α that the interest rate will go up by 1%, and 1 minute ago a very unreliable agent advises α that the interest rate may go up by 0.5%, then the second unreliable chunk should not be permitted to 'overwrite' the first. We capture this by only permitting a new chunk of *info* to be imported if the resulting distribution has more information *relative to* the decay limit distribution than the existing distribution has. Precisely, this is measured using the Kullback-Leibler distance measure[5], and *info* is only used if:

$$\sum^n_{i=1} p^t_i \log\frac{p^t_i}{d^D_i} > \sum^n_{i=1} q^t_i \log\frac{q^t_i}{d^D_i} \tag{8}$$

In addition, we have described in Eqn. 5 how the integrity of each distribution D will decay in time. Combining these two into one result, distribution D is revised to:

$$q^{t+1}_i = \begin{cases} (1-\nu^D)\times d^D_i + \nu^D\times p^t_i & \textit{if } \text{usable } \mathit{info} \text{ is received at time } t \\ (1-\nu^D)\times d^D_i + \nu^D\times q^t_i & \textit{otherwise}\end{cases}$$

for $i = 1,\cdots,n$, and decay rate ν^D as before. We have yet to estimate $R^t(\alpha,\beta,O(\mathit{info}))$ — that is described in Sec. 6.2. We now illustrate the above by showing how an important class of *info* illocutions, namely preference information, is managed.

Preferences. Preference information is a statement by an agent that it prefers one class of deals to another where deals may be multi-issue. Preference illocutions may refer to particular issues within deals — e.g. "I prefer red to yellow", or to combinations of issues — e.g. "I prefer a car with a five year warranty to the same car with a two year warranty than costs 15% less". Here α receives preference information, *info*, from β through an Inform$(\beta,\alpha,\mathit{info})$ illocution, or through one of the other argumentation illocutions — see Sec. 2 — and attaches $R^t(\alpha,\beta,O(\mathit{info}))$ to it.

[5]This is just one criterion for determining whether the *info* should be used.

What happens next will depend on α's plans. Suppose that α has a plan s that constructs the probability distribution $P^t(\text{UPrefer}(\beta, \alpha, \delta))$ over all δ meaning that "δ is the deal that β prefers most". α may know the decay limit distribution for $\text{UPrefer}(\cdot)$. Suppose α has a prior distribution $(q_i^t)_{i=1}^n$ for $\text{UPrefer}(\cdot)$. Now suppose that α learns the *info* "x% of the time β prefers deals with property Q_1 to those with property Q_2". J_s^{UPrefer} derives the following linear constraint on the $P^t(\text{UPrefer}(\beta, \alpha, \delta))$ distribution:

$$\frac{x}{100} = \frac{\sum_{\delta:Q_1(\delta)} p_\delta}{\left(\sum_{\delta:Q_1(\delta)} p_\delta\right) + \left(\sum_{\delta:Q_2(\delta)} p_\delta\right) - \left(\sum_{\delta:Q_1 \wedge Q_2(\delta)} p_\delta\right)}$$

and the procedure described above follows as described. Note that this manages two different probabilities: first, the probability within the *info* "x% of the time...", and second α's estimate of β's reliability in providing information of this sort, $R^t(\alpha, \beta, O(\textit{info}))$, that is determined by the ontological context $O(\textit{info})$ of the *info*.

6.2 A reliability model

We estimate $R^t(\alpha, \beta, O(\textit{info}))$ by measuring the error in information. α's plans will have constructed a set of distributions. We measure the 'error' in information as the error in the effect that information has on each of α's distributions. Suppose that a chunk of *info* is received from agent β at time u and is verified at some later time t. For example, a chunk of information could be "the interest rate will rise by 0.5% next week", and suppose that the interest rate actually rises by 0.25% — call that correct information *fact*. What does all this tell agent α about agent β's reliability? Consider one of α's distributions D that is $\{q_i^u\}$ at time u. Let $(p_i^{\textit{info}})_{i=1}^n$ be the minimum relative entropy distribution given that *info* has been received as calculated in Eqn. 6, and let $(p_i^{\textit{fact}})_{i=1}^n$ be that distribution if *fact* had been received instead. Suppose that the reliability estimate for distribution D was R_D^u. This section is concerned with what R_D^u should have been in the light of knowing *now*, at time t, that *info* should have been *fact*, and how that knowledge effects our current reliability estimate for D, $R^t(\alpha, \beta, O(\textit{info}))$.

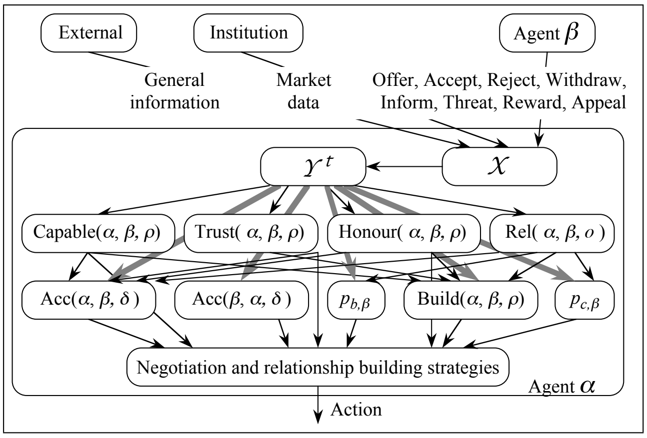
The idea of Eqn. 7, is that the current value of r should be such that, *on average*, $(p_i^u)_{i=1}^n$ will be seen to be "close to" $(p_i^{\textit{fact}})_{i=1}^n$ when we eventually discover *fact* — no matter whether or not *info* was used to update D, as determined by the acceptability test in Eqn. 8 at time u. That is, given *info*, *fact* and the prior $(q_i^u)_{i=1}^n$, calculate $(p_i^{\textit{info}})_{i=1}^n$ and $(p_i^{\textit{fact}})_{i=1}^n$ using Eqn. 6. Then the *observed reliability* for distribution D, $R_D^{(\textit{info}|\textit{fact})}$, on the basis of the verification of *info* with *fact* is the value of r that minimises the Kullback-Leibler distance between $(p_i^u)_{i=1}^n$ and $(p_i^{\textit{fact}})_{i=1}^n$:

$$\arg\min_r \sum_{i=1}^n (r \cdot p_i^{\textit{info}} + (1-r) \cdot q_i^u) \log \frac{r \cdot p_i^{\textit{info}} + (1-r) \cdot q_i^u}{p_i^{\textit{fact}}}$$

If $E^{\textit{info}}$ is the set of distributions that *info* affects, then the overall *observed reliability* on the basis of the verification of *info* with *fact* is: $R^{(\textit{info}|\textit{fact})} = 1 - (\max_{D \in E^{\textit{info}}} |1 - R_D^{(\textit{info}|\textit{fact})}|)$. Then for each ontological context o_j, at time t when, perhaps, a chunk of *info*, with $O(\textit{info}) = o_k$, may have been verified with *fact*:

$$R^{t+1}(\alpha, \beta, o_j) =$$
$$(1-\nu) \times R^t(\alpha, \beta, o_j) + \nu \times R^{(\textit{info}|\textit{fact})} \times \text{Sim}(o_j, o_k)$$

Figure 1: Trust, Honour and Reliability in the informed agent. The figure shows the social model and a sample world model. The social model consists of the four components: $\text{Capable}(\cdot)$, $\text{Trust}(\cdot)$, $\text{Honour}(\cdot)$ **and** $\text{Rel}(\cdot)$ **standing for 'reliability'. The sample world model contains just five distributions with the thick arcs representing the five model building functions** J_s^D**.**



where Sim measures the semantic distance between two sections of the ontology, and ν is the learning rate. Over time, α notes the ontological context of the various chunks of *info* received from β and over the various ontological contexts calculates the relative frequency, $P^t(o_j)$, of these contexts, $o_j = O(\textit{info})$. This leads to a overall expectation of the *reliability* that agent α has for agent β:

$$R^t(\alpha, \beta) = \sum_j P^t(o_j) \times R^t(\alpha, \beta, o_j)$$

7. AGENT ARCHITECTURE

The agent architecture developed so far is summarised in Fig. 1. α's actions are determined by its relationship-building strategy that identifies trading partners, and by its negotiation strategy that manages individual negotiations. The complete apparatus in Fig. 1 is revised when any illocution is received.

The *social model*: $\text{Capable}(\cdot)$, $\text{Trust}(\cdot)$, $\text{Honour}(\cdot)$, and $\text{Reliability}(\cdot)$, presents a summary of the interaction history weighted towards recent observations. It provides the foundation for α's relationship building strategy. The estimation of $\text{Capable}(\cdot)$ depends on the commitments involved — for example, estimating whether β is capable of raising a sum of money presents a different problem to estimating whether β can supply 100 kilos of sausages — the estimation of $\text{Capable}(\cdot)$ is not discussed here. An estimation of $\text{Trust}(\cdot)$ is discussed in [14], and herein Honour in Sec. 5 and Reliability in Sec. 6. The world model is described following.

7.1 World Model

Agent α acts by activating a plan s that invokes model building functions J_s^D and uses one of the two entropy inference methods [see Sec. 4] to "fill in the gaps" in probability distribution D. The *world model* is then aggregated from these distributions. Fig. 1 shows a sample world model consisting of: $\text{Acc}(\alpha, \beta, \delta)$, $\text{Acc}(\beta, \alpha, \delta)$, $p_{b,\beta}$, $\text{Build}(\alpha, \beta, \rho)$ and $p_{c,\beta}$. The components of the world model estimate of α's belief that a proposition is true.

$P^t(\text{Acc}(\alpha, \beta, \delta))$ is α's estimate that δ is an acceptable deal from agent β — this probability expresses the certainty that α has in this proposition. So $P^t(\text{Acc}(\alpha, \beta, \delta_1)) > P^t(\text{Acc}(\alpha, \beta, \delta_2))$ does not mean that δ_1 is an intrinsically better deal than δ_2 — it means that α is more certain that δ_1 is acceptable than δ_2 is acceptable.[6] A framework for estimating the $P^t(\text{Acc}(\alpha, \beta, \delta))$ distribution, derived from subjective, market and general observations, is described in [1].

$P^t(\text{Acc}(\beta, \alpha, \delta))$ is α's estimate that δ is an acceptable deal to β. [1] estimates $P^t(\text{Acc}(\beta, \alpha, \delta))$ by observing the $\text{Offer}(\cdot)$, $\text{Accept}(\cdot)$ and $\text{Reject}(\cdot)$ illocutions received from β.

$p_{b,\beta}$ is estimated by observing β's behaviour. A simple method is to measure the local entropy of the $P^t(\text{Acc}(\beta, \alpha, \delta))$ distribution — that is, the entropy of the normalisation of that distribution for δ restricted to the negotiation region. If that entropy measurement does not fall in time as the negotiation proceeds then this may indicate β's intention to withdraw from the negotiation. [1] describes the estimation of $p_{b,\beta}$.

A natural way to define $\text{Build}(\alpha, \beta, \rho)$ is by combining the four concepts we claimed were its pillars:

$$P(\text{Build}(\alpha, \beta, \rho)) = k(\text{Capable}(\alpha, \beta, \rho),$$
$$\text{Trust}(\alpha, \beta, \rho), \text{Honour}(\alpha, \beta, \rho), \text{Rel}(\alpha, \beta, O(\rho)))$$

for some function k, where $O(\rho)$ denotes the ontological context of relationship ρ. This characterisation of Build is defined in terms of (commitment, observation) pairs for each of its four components.

A simple negotiation agent aims to strike a deal δ with agent β whilst avoiding the risk of negotiation breakdown. Similarly, an argumentation agent aims to build a trading relationship ρ with agent β whilst avoiding the risk of relationship collapse. To estimate $p_{c,\beta}$ we observe β for odd behaviour. Indications of collapse include: individual negotiations that have a high number of rounds or that take a significant time, reluctance to benefit from attractive terms offered in α's $\text{Reward}(\cdot)$s, β offers unrealistic $\text{Reward}(\cdot)$s, β issues $\text{Threats}(\cdot)$s, or, if the structure of the electronic institution permits general observation of β's trading, that β has taken his business elsewhere.

8. CONCLUSION

We introduced in this paper two new concepts: $\text{Honour}(\cdot)$ and $\text{Reliability}(\cdot)$. $\text{Honour}(\cdot)$ models the fulfilment of the promises derived from the use of argumentative particles such as *appeals*, *rewards* and *threats*. This concept plays a similar role to that that the concept of $\text{Trust}(\cdot)$ plays with respect to the fulfilment of contract agreements. $\text{Reliability}(\cdot)$ models the reliability of the information tabled by an opponent. We have proposed a semantics for these two concepts based on information theory. Also, an agent architecture to support this model is proposed.

[6] We differ from game-theory-based agents in that our agent is not necessarily "utility aware".

9. REFERENCES

[1] Debenham, J. Bargaining with information. In *Proceedings Third International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2004* (July 2004), N. Jennings, C. Sierra, L. Sonenberg, and M. Tambe, Eds., ACM, pp. 664 – 671.

[2] Dorigo, M., and Stützle, T. *Ant Colony Optimization.* MIT Press, Cambridge, MA, 2004.

[3] Esteva, M., Padget, J., and Sierra, C. Formalizing a languages for institutions and norms. In *Intelligent Agents VIII*, J. Meyer and M. Tambe, Eds. Springer-Verlag, Berlin, Germany, 2002, pp. 348 – 366.

[4] Faratin, P., Sierra, C., and Jennings, N. Using similarity criteria to make issue trade-offs in automated negotiation. *Journal of Artificial Intelligence 142*, 2 (2003), 205–237.

[5] Giménez, E., Godo, L., Rodríguez-Aguilar, J. A., and Garcia, P. Designing bidding strategies for trading agents in electronic auctions. In *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS-98)* (1998), pp. 136–143.

[6] Halpern, J. *Reasoning about Uncertainty.* MIT Press, 2003.

[7] Jaynes, E. *Probability Theory — The Logic of Science.* Cambridge University Press, 2003.

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

[9] Kraus, S. Negotiation and cooperation in multi-agent environments. *Artificial Intelligence 94*, 1–2 (1997), 79–97.

[10] Li, Y., Bandar, Z. A., and McLean, D. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering 14(4)* (2003).

[11] MacKay, D. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, 2003.

[12] Paris, J. Common sense and maximum entropy. *Synthese 117*, 1 (1999), 75 – 93.

[13] Rosenschein, J. S., and Zlotkin, G. *Rules of Encounter.* The MIT Press, Cambridge, USA, 1994.

[14] Sierra, C., and Debenham, J. An information-based model for trust. In *Proceedings Fourth International Conference on Autonomous Agents and Multi Agent Systems AAMAS-2005* (Utrecht, The Netherlands, July 2005), F. Dignum, V. Dignum, S. Koenig, S. Kraus, M. Singh, and M. Wooldridge, Eds., ACM Press, New York, pp. 497 – 504.

[15] Sierra, C., Jennings, N., Noriega, P., and Parsons, S. *Proceedings of the 4th International Workshop on Intelligent Agents IV, Agent Theories, Architectures, and Languages.* Springer-Verlag, London, UK, 1997, ch. A Framework for Argumentation-Based Negotiation, pp. 177 – 192.