# Stereotypical Trust and Bias in Dynamic Multiagent Systems

CHRIS BURNETT and TIMOTHY J. NORMAN, University of Aberdeen
KATIA SYCARA, Carnegie Mellon University

Large-scale multiagent systems have the potential to be highly dynamic. Trust and reputation are crucial concepts in these environments, as it may be necessary for agents to rely on their peers to perform as expected, and learn to avoid untrustworthy partners. However, aspects of highly dynamic systems introduce issues which make the formation of trust relationships difficult. For example, they may be short-lived, precluding agents from gaining the necessary experiences to make an accurate trust evaluation. This article describes a new approach, inspired by theories of human organizational behavior, whereby agents generalize their experiences with previously encountered partners as *stereotypes*, based on the observable *features* of those partners and their behaviors. Subsequently, these stereotypes are applied when evaluating new and unknown partners. Furthermore, these stereotypical opinions can be communicated within the society, resulting in the notion of *stereotypical reputation*. We show how this approach can complement existing state-of-the-art trust models, and enhance the confidence in the evaluations that can be made about trustees when direct and reputational information is lacking or limited. Furthermore, we show how a stereotyping approach can help agents detect unwanted biases in the reputational opinions they receive from others in the society.

Categories and Subject Descriptors: I.2.11 [**Distributed Artificial Intelligence**]: Multi-Agent Systems

General Terms: Reliability, Experimentation, Design

Additional Key Words and Phrases: Trust, multiagent systems, stereotypes

**ACM Reference Format:**
Burnett, C., Norman, T. J., and Sycara, K. 2013. Stereotypical trust and bias in dynamic multiagent systems. ACM Trans. Intell. Syst. Technol. 4, 2, Article 26 (March 2013), 22 pages.
DOI = 10.1145/2438653.2438661 http://doi.acm.org/10.1145/2438653.2438661

## 1. INTRODUCTION

Trust is a vital concept in open and dynamic MultiAgent Systems (MAS), where diverse agents continually join, interact, and leave. In such environments, some agents will inevitably be more trustworthy than others, displaying varying degrees of competence and self-interest in different interactions. When faced with the problem of choosing a partner with whom to interact, agents should evaluate the potential candidates and determine which one is the most appropriate with respect to a given interaction and context. When making such evaluations, trust plays an important role. Trust is a rich

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.
Authors' addresses: C. Burnett (corresponding author) and T. J. Norman, Computing Science Department, Meston Building, King's College, University of Aberdeen, AB24 3FX, UK; email: cburnett@abdn.ac.uk; K. Sycara, Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213.

DOI 10.1145/2438653.2438661 http://doi.acm.org/10.1145/2438653.2438661

concept, and can be defined and modeled in different ways, and to different levels of granularity. While it has been well argued that trust should be represented as a rich cognitive structure of beliefs [Castelfranchi and Falcone 1998], we define trust here pragmatically as the *degree of belief*, according to the definition of Gambetta [1990].

> Trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action... When we say we trust someone or that someone is trustworthy, we implicitly mean that the probability that he will perform an action that is beneficial or at least not detrimental to us is high enough for us to consider engaging in some form of cooperation with him. [Gambetta 1990]

By taking this view, we can show how the work in this article may be applicable to the extensive trust literature which shares this *probabilistic* view of trust [Burnett et al. 2010; Huynh et al. 2006; Jøsang and Ismail 2002; Teacy et al. 2006; Wang and Singh 2007].

State-of-the-art trust approaches generally consider an agent's trust in a potential partner as a function of the evidence available about that partner, whether they are directly experienced, relayed by other agents in the society, or produced by some static organizational rules [Huynh et al. 2006]. If an agent has insufficient direct evidence to form a confident evaluation of another, it can make use of *reputational* [Sabater 2003] evidence by obtaining the opinions of other agents who have previously interacted with it. In highly dynamic societies, however, the formation and maintenance of trust relationships with these methods can be difficult. We characterize such societies as those with the following features.

—*Diverse and self-interested*. Agents may pursue their own goals, and display varying levels of competence and trustworthiness.
—*Dynamicity*. Agents may join and leave the society with high frequency.
—*Ad hoc organizational structures*. Agents may be assembled into short-term ad hoc groups (such as coalitions) to achieve a particular shared goal. Furthermore, interaction may be constrained to within these groups.

In such conditions, agents may be precluded from gathering sufficient experiences from partners to form stable trust relationships. As a result, both direct and reputational evaluations may be sparse or unavailable. Similarly, agents may frequently be prevented from interacting with trusted partners, either due to the shifting organizational structures in the society, or the high rate of agent turnover within the population. Under these conditions, the utility of traditional trust approaches may be severely limited.

We present in this article a model of *stereotypical* trust which aims to address these issues. Through stereotyping, agents generalize their experiences with known partners in previous contexts in order to form stereotypical evaluations about unknown agents in new contexts. By ascribing trust evaluations to learned *classes* of individuals as well as individuals themselves, agents can make use of previous experiences and reputational opinions in contexts where this would not otherwise be possible. We show how this approach can provide benefits when there exist correlations between the behaviors of agents, and the features they possess. We also show how a stereotyping approach can be applied to the problem of selecting appropriate reputation providers when stereotypical biases exist in the society.

The remainder of the article proceeds as follows. In Section 2 we present our rationale for stereotypical trust evaluations. In Section 3 we provide an overview of the trust evaluation framework we have adopted. In Section 4 we present our stereotypical

trust model. In Section 5 we discuss the problem of stereotypical biases, and present a mechanism for mitigating the effects of such biases. In Section 6 we evaluate the performance of our model in highly dynamic environments. Finally, we present a discussion of our findings and related work in Section 7 and conclude in Section 8.

## 2. STEREOTYPES AND TRUST

While ad hoc environments can be problematic for trust models in multiagent systems, a number of authors have found that this is not always the case in highly dynamic human organizations [Jarvenpaa and Leidner 1999; Meyerson et al. 1996], such as multiagency emergency response teams [Militello et al. 2007; Carver and Turoff 2007]. When diverse individuals that are unfamiliar with each other form teams to solve problems or cooperate, some initial and tentative form of trust has been found to be present. The theory of *swift trust* was developed by Meyerson et al. [1996] to characterize this trust, and the processes responsible for its formation. The authors studied the trust relationships that form within film studio crews, who are assembled from a variety of diverse organizations, and often work together for the duration of one project before disbanding. The authors found that crew members behaved as if some initial form of trust was present, despite a lack of direct or reputational opinions to draw from. A key source of this initial trust is the process by which individuals generalize their experiences with others to *categorical* experiences, where categories are defined by the presence (or absence) of certain salient features.

While the notion of stereotyping often carries a negative connotation, stereotypes may represent an accurate reflection of reality [Hilton and Von Hippel 1996], based on rational generalizations from personal experiences. For example, when an employer seeks to hire a new employee, he will likely produce a "short-list" of candidates based on their CVs (curriculum vitae). This task involves assessing an individual solely on the basis of the "features" visible in the CV documents. Educational qualifications will be considered, as may be previous places of employment, hobbies, and so on. From this, the employer will use his accumulated knowledge of relationships between visible features and trustworthiness to make a tentative evaluation of the candidate, and a subsequent decision, based on these features.

While we will generally discuss stereotypes and features in the abstract in this article, it is worth mentioning here some of the possible sources of feature information that may be available within a multiagent system. In the most intuitive sense, we may consider the visible attributes of agents as features. For example, with software agents, observable attributes can include the trustee agent's owner, programmer, user, or version number. If agents represent human users of a system, as may be the case in e-commerce applications, features may include attributes such as nationality, location, age, and so on. Furthermore, features may be "bestowed" by other agents or institutions. For example, accreditations or certificates may be obtained from accreditation institutions (for example, Certification Authorities [Eschenauer et al. 2003]) used to indicate that an individual is competent or trustworthy according to those institutions.

However, with some additional reasoning, it may be possible to obtain features from other observations. For example, observations about an agent's accumulated experience in different tasks can be easily converted into features. For example, we may create a "feature" which represents every ten instances of a task performed by an agent, creating experience "milestones". If an agent has performed a task $\tau$ 23 times, we may signify this with the feature "$n(\tau) \geq 20$", meaning "performed $\tau$ at least 20 times". By interacting with agents with different levels of experience, trustors can build stereotyping models based on these experiential features, which can then be used to predict the trustworthiness of new, unknown agents. These models then

represent the notion of expected "learning curves" for tasks, as they estimate the trustworthiness of agents given the experience they have accumulated.

If agents are situated within a social network [Jøsang et al. 2006; Hang et al. 2009], then observable social relationships may also provide a useful source of feature information. As features are simply binary variables in our model, the feature representing a relationship of type $R$ from one agent, $a$, to another, $b$, could be represented simply by $a$ possessing the feature $Rb$, and $b$ possessing the feature $aR$. Bidirectional social relationships can be represented as two features, one for each direction of the relationship.

In the remainder of this article, we will present our approach whereby agents interacting in MAS can make use of available features in order to make trust evaluations when evidence is unavailable. Agents can build stereotypes which attempt to model their observations as accurately as possible. Where relationships exist between agent features and behavior, we will show that a stereotyping approach can help agents to avoid the need to engage in exploration or random partner selection.

## 3. FRAMEWORK

We define here the common framework that we will use throughout to describe agents in a multiagent society, and the tasks they can perform. We will introduce the underlying trust evaluation model that we will use to evaluate our stereotyping approach, which will be presented in the following section (Section 4).

### 3.1. Agents and Tasks

We assume a society of agents, $A = \{x, y \dots\}$, which we refer to as the *global* society. Where we are concerned with the specific role of the agent, we use lowercase $x$ to represent some agent $x \in A$ playing the role of a *trustor*, and lowercase $y$ to represent some agent $y \in A$ playing the role of a *trustee*. As we consider trust to be specific to a particular issue, we assume a set $\mathcal{T} = \{\tau_1 \dots \tau_n\}$ of possible tasks[1]. Each task $\tau \in \mathcal{T}$ has a number of possible outcomes $\mathcal{O}_\tau$. In order to determine whether a particular outcome $o_\tau$ represents success or failure (or satisfaction/dissatisfaction), agents make use of their own *subjective evaluation function* for that task. This function represents the notion that different agents may have different expectations about what constitutes good or bad performance in a given task. The subjective evaluation function of an agent $x$ for a task $\tau$ is denoted $\varsigma_\tau^x : \mathcal{O}_\tau \to \{0, 1\}$, where 0 represents task failure, and 1 represents success. Each agent can then partition the set of possible task outcomes into those that, in its opinion, represent success and failure, denoted as $O_{x:\tau}^+$ and $O_{x:\tau}^-$ respectively, such that $O_{x:\tau}^+ = \{o | o \in \mathcal{O}_\tau \wedge \varsigma_\tau^x(o) = 1\}$, $O_{x:\tau}^- = \{o | o \in \mathcal{O}_\tau \wedge \varsigma_\tau^x(o) = 0\}$, $O_{x:\tau}^+ \cup O_{x:\tau}^- = \mathcal{O}_\tau$ and $O_{x:\tau}^+ \cap O_{x:\tau}^- = \emptyset$.

As we are interested in ad hoc team situations, we partition the global population into a number of *ad hoc groups*, denoted as $\mathcal{G} = \{G_1 \dots G_n\}$, $\forall G \in \mathcal{G}, G \subset A$, analogous to coalitions or teams. It is not necessary for the set of ad hoc groups to completely cover the global population. We require, however, that agents be in only one group at any given moment, such that $\forall G, G' \in \mathcal{G}, (G \neq G' \to G \cap G' = \emptyset)$.

We denote the group to which a particular agent $x$ belongs as $G_x$. In addition, we define $R_x \subset A$ to be the set of recommender agents visible to $x$, and $Y_x \subset A$ to be the set of candidate trustees visible to $x$. Agents are visible to each other if they are in the same ad hoc group, so we define these sets for each agent as $\forall x \in A, R_x = G_x \setminus \{x\}$, and $\forall x \in A, Y_x = G_x$. Note that our definition of $Y_x$ allows trustors to consider delegating to themselves. This conforms to the view of Castelfranchi and Falcone [1998], whose cognitive model of trust allows agents to form beliefs about their *own* trustworthiness

[1]While we refer to *tasks* in this article, these could represent any issue an agent can form an expectation about, such as adherence to a social norm, or provision of reliable evidence within a subject area.

with respect to a particular issue. In this way, an agent can consider itself as a potential candidate for delegation, and will only consider delegating tasks to trustees more trusted than itself.

Each trustor $x \in A$ maintains a set of *opinions* $Ops_x$ about previously encountered trustees, and an *experience base* $Eb_x$ of past experiences (i.e., directly observed interaction outcomes) with those trustees, from which opinions can be formed. The particular opinion representation we adopt is outlined in detail in Section 3.2. Each agent $x \in A$ also possesses a trust evaluation function $E_x(y, \tau, Ops_x, R_x, S_x)$ which returns a degree of trust for a trustee $y \in A$ with respect to a task $\tau \in \mathcal{T}$, given the experience base of $x$, $Eb_x$, the set of existing opinions held by $x$, $Ops_x$, and those of the visible recommenders, $R_x$. The *stereotypical* evaluations produced by the stereotyping model $S_x$ also influence the evaluation function. The behavior of the function $E_x$ is described in the course of this section, while the function $S_x$ is described in Section 4.

We assume a finite set of binary *features* $\mathcal{F} = \{f_1 \dots f_n\}$ which defines the set of all possible features agents may possess. Each agent $x \in A$ has a visible *feature vector* containing values for some or all of the possible features and denoted by $F_x \subseteq \mathcal{F}$. Since features in our model are binary variables, a feature vector $F_x$ of an agent $x$ can say one of three things about a particular feature $f \in \mathcal{F}$. If $f \in F_x$, we can say that $x$ has feature $f$. If $\neg f \in F_x$, then we can say that $x$ does not have that feature. However, if some features in $\mathcal{F}$ are not in an agent's feature vector, we cannot say anything about their presence, and refer to them as *unobserved* features. The handling of such features is described in Section 4.

While we discuss features here in the abstract, they may come from a variety of sources within a real multiagent system.

### 3.2. Representing Opinions

The aim of trust evaluation models is to produce subjective beliefs, or *opinions*, about the trustworthiness of individuals with respect to different issues, based on evidence. While our approach does not restrict the choice of trust opinion representation, we present here a simple trust evaluation model based on Subjective Logic (SL) [Jøsang et al. 2007], as its relatively simple notation allows for an intuitive discussion of the integration of stereotypical evaluations with the traditional sources of direct and reputational evidence.

*3.2.1. Belief Representation.* An opinion held by an agent $x$ about agent $y$ performing a task $\tau$ is represented as a tuple

$$\omega^x_{y:\tau} = \left\langle b^x_{y:\tau}, d^x_{y:\tau}, u^x_{y:\tau}, a^x_{y:\tau} \right\rangle \tag{1}$$

$$\text{where } b^x_{y:\tau} + d^x_{y:\tau} + u^x_{y:\tau} = 1, \tag{2}$$

$$\text{and } a^x_{y:\tau} \in [0, 1]. \tag{3}$$

In the preceding opinion representation, $b^x_{y:\tau}, d^x_{y:\tau}, u^x_{y:\tau}, a^x_{y:\tau}$ represent the degrees of belief, disbelief, uncertainty, and the base rate (or a priori degree of belief), respectively. In the context of trust, we use the term belief to mean the extent to which $x$ believes that delegating $\tau$ to $y$ will result in a positive outcome. The $u$ parameter represents the uncertainty about the probability of an event, and as such represents a kind of second-order uncertainty[2]. In each case, the superscript identifies the belief *owner*, and the subscript represents the belief *target*, that is, the agent and task that the opinion pertains to.

[2]We will use the term *ambiguity* to refer to the type of uncertainty represented by the $u$ parameter to avoid confusion with the uncertainty inherent in the opinion as a whole, since it itself represents a probability distribution.

*3.2.2. Evidence Aggregation.* Agents base their opinions on evidence, which is obtained by interacting with, and subsequently evaluating, other agents. Alternatively, evidence can be obtained from third parties who have interacted with a particular individual before. Opinions in SL are formed by aggregating positive and negative evidence about a particular individual. A body of evidence held by an agent $x$ about another $y$ is a pair $\langle r^x_{y:\tau}, s^x_{y:\tau}\rangle$, where $r^x_{y:\tau}$ is the number of positive experiences observed by $x$ about $y$, and $s^x_{y:\tau}$ is the number of observed negative experiences. Eqs. (4), (5), and (6) show how the $r^x_{y:\tau}$ and $s^x_{y:\tau}$ parameters are used to produce an opinion [Jøsang et al. 2006]

$$b^x_{y:\tau} = \frac{r^x_{y:\tau}}{r^x_{y:\tau} + s^x_{y:\tau} + 2} \tag{4}$$

$$d^x_{y:\tau} = \frac{s^x_{y:\tau}}{r^x_{y:\tau} + s^x_{y:\tau} + 2} \tag{5}$$

$$u^x_{y:\tau} = \frac{2}{\left(r^x_{y:\tau} + s^x_{y:\tau} + 2\right)} \tag{6}$$

These equations provide a simple mapping from observed evidence about an agent in a particular task, to an opinion in the opinion representation given before. Eq. (6) ensures that uncertainty decreases as more evidence is observed. We note, however, that more sophisticated mappings exist. Wang and Singh [2007] present a method for updating trust opinions with a formulation of uncertainty which is sensitive to both the quantity of evidence observed, and the conflict within that evidence. However, the model presented here is illustrative of probabilistic trust models, and is sufficient to demonstrate our stereotyping approach, discussed in the following section.

*3.2.3. Probability Expectation Value.* An opinion's probability expectation value can be used as a single-valued trust metric, suitable for ranking potential partners. Eq. (7) shows how a probability expectation value $P(\omega^x_{y:\tau})$ is calculated from an opinion $\omega^x_{y:\tau}$. We use the term *rating* to mean $P(\omega^x_{y:\tau})$ for a particular opinion $\omega^x_{y:\tau}$.

$$P\left(\omega^x_{y:\tau}\right) = b^x_{y:\tau} + a^x_{y:\tau} \cdot u^x_{y:\tau} \tag{7}$$

By using Eq. (7), (4), and (6) together, we can obtain, for a given evidence pair $\langle r^x_{y:\tau}, s^x_{y:\tau}\rangle$, a probability expectation value $P(\omega^x_{y:\tau})$.

*3.2.4. Base Rate.* The base rate parameter $a^x_{y:\tau}$ represents the a priori degree of trust $x$ has about $y$ performing task $\tau$, before any evidence has been received. It determines the effect that the parameter $u^x_{y:\tau}$ will have on the resultant probability expectation value. In order to convert an opinion in belief representation to an opinion (or *rating*) compatible with classical probability theory, it is necessary to "resolve" the unassigned ambiguity MAS in some way. This involves transferring the free ambiguity belief MAS in an opinion to either belief or disbelief. As can be seen from Eq. (7), this transfer is dependant on the base rate parameter $a^x_{y:\tau}$. The default value of $a^x_{y:\tau}$ is the uninformative prior 0.5, which means that before any positive or negative evidence has been received, outcomes from both $O^+_{a:\tau}$ and $O^-_{a:\tau}$ are considered equally likely. In this case, $P(\omega^x_{y:\tau}) = 0.5$, which is the least informative value it can take. Values of $a^x_{y:\tau} > 0.5$ will result in more ambiguity MAS being converted to belief, and conversely disbelief for $a^x_{y:\tau} < 0.5$.

Figure 1 shows an example set of opinions. Due to the additivity requirement of the $b^x_{y:\tau}$, $d^x_{y:\tau}$ and $u^x_{y:\tau}$ parameters, there are only two degrees of freedom when plotting opinions. This means that we do not need to represent opinions in three dimensions. The opinion spaces of agents can be visualized as a triangular (ternary) plot, with the

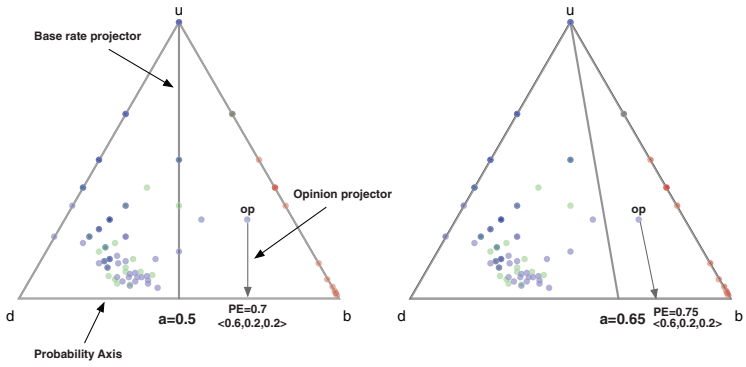



Fig. 1. Sample opinion space for a trustor, with base rate projector shown.

top vertex representing maximum ambiguity, the bottom left representing maximum disbelief, and the bottom right representing maximum belief. The distance from the midpoint of the leftmost edge represents the degree of belief, the distance from the midpoint of the rightmost edge represents disbelief, and the distance from the bottom edge represents the degree of ambiguity in the opinion. This representation provides a helpful tool to visualize the effect of stereotypical base rates on the $P(\omega)$ values produced by the model.

The bottom edge of the triangle represents the classical probability axis. Opinions lying on this edge are considered to be *dogmatic*, in that they contain no ambiguity. The base rate value $a^x_{y:\tau}$ is plotted along this edge. In calculating $P(\omega^x_{y:\tau})$, opinions are *projected* onto this axis following a line parallel to the base rate projector line (originating at the top vertex and ending at the point marked "$a$" on the probability axis). Figure 1 shows an example opinion space with two different base rates. The leftmost opinion has $a^x_{y:\tau} = 0.5$, representing an opinion with no stereotypical component. The rightmost opinion has $a^x_{y:\tau} = 0.65$. As a result, any unassigned ambiguity MAS in the opinion will resolve more favorably than in the leftmost opinion. The probability expectation value for the example opinion is then shifted from 0.7 to 0.75. If this opinion was entirely unsupported by evidence, the value of $u^x_{y:\tau}$ would be 1, and the resulting probability expectation would be $P(\omega^x_{y:\tau}) = 0.65$. If these two opinions relate to the same trustee, then the rightmost opinion represents a more optimistically biased view, even if the opinions are based on the same evidence parameters.

*3.2.5. Reputation.* Reputation in probabilistic trust systems is often calculated by aggregating the $r^x_{y:\tau}$ and $s^x_{y:\tau}$ parameters from reputation providers [Wang and Singh 2007]. The result of the aggregation of evidence provided by a set of recommender agents $R$ is a combined evidence pair $\langle r'^x_{y:\tau}, s'^x_{y:\tau} \rangle$.

$$r'^x_{y:\tau} = r^x_{y:\tau} + \sum_{\rho \in R_x} r^\rho_{y:\tau} \qquad s'^x_{y:\tau} = s^x_{y:\tau} + \sum_{\rho \in R_x} s^\rho_{y:\tau} \tag{8}$$

Once the evidence parameters have been aggregated, an opinion and rating for the combined evidence can be calculated using Eqs. (4) and (7). We note that a strong assumption is made here; it may be unwise to aggregate the opinions of all available providers, as some may be unreliable, malicious, or biased in some way. It is therefore important to consider the trust in opinion sources, by discounting opinions from these sources. While addressing such issues remains an open problem, a number of approaches exist. These include approaches based on transitive trust networks [Jøsang

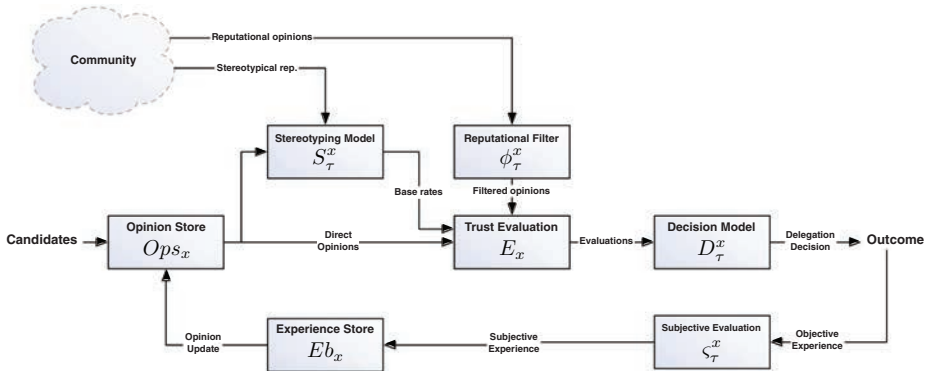



Fig. 2. Overview of the stereotypical trust evaluation framework.

et al. 2006; Hang et al. 2009], reputational filtering mechanisms [Sensoy et al. 2009; Teacy et al. 2006], and logical representations of graded trust [Lorini and Demolombe 2008]. However, in order to clearly present our contributions, and without loss of generality, we assume a simple model of reputation here. In Section 5, we present a *stereotypical* reputational filtering mechanism, which attempts to address this problem when the biases of opinion providers correlate with the features of agents.

### 3.3. Decision Model

The process of evaluating potential partners is distinct from that of deciding which partner to choose, and whether to delegate at all. To permit a clear discussion and evaluation of our stereotyping approach, we assume a very simple trust decision model in this article. Given a number of potential candidates $Y_x$, a trustor $x$ will always select the highest rated candidate in $Y_x$ for a task $\tau$, denoted $C_{x:\tau}$, according to that trustor's evaluation model, such that $C_{x:\tau} = \arg\max_{y \in Y_x} E_x(y, \tau, Ops_x, R_x)$. While this decision model permits a clear investigation of the merits of our stereotyping approach, it is important to note that it is too simple for practical use. Castelfranchi and Falcone [1998] argue that social decision-making requires the integration of trust assessments with the rewards, risks, and contexts inherent in a particular situation. Our approach does not preclude a more sophisticated decision-making approach, and future work will investigate techniques for improving the quality of trust decision-making in highly dynamic systems, as well as trust evaluation.

### 3.4. Summary

Figure 2 provides an overview of how these components fit together to form an agent's trust evaluation and decision mechanism. When evaluating a set of candidates, an agent $x$ first retrieves its own opinions about those candidates from its opinion store $Ops_x$. These opinions are then passed to the trust evaluation function $E_x$. At this point the community of reputation providers may also be queried for third-party opinions. Together with direct opinions, these are used by the evaluation function to produce a set of evaluations. Using the simple decision model $C_{x:\tau}$, the most trusted agent is selected for delegation. When the delegation is completed, the agent $x$ observes the outcome, and evaluates it using its own (task-specific) subjective evaluation function $\varsigma^x_\tau$ to produce a subjective outcome. This new outcome is used to update the opinion of $x$ about the selected candidate, and the process may begin anew. In the following section, we introduce the *stereotyping model* $S^x_\tau$, one of the key contributions of this article, which produces a priori stereotypical evaluations which can assist the evaluation function

when limited direct or reputational opinions are available. In Section 5, we discuss the *reputational filtering* component of the architecture ($\phi_\tau^x$), which addresses the problem of stereotypical biases within the community of opinion providers.

Note that we do not "feed back" stereotypically biased opinions into the opinion base. This can result in a kind of "confirmation bias" [Chen and Bargh 1997], where new evidence is interpreted in a biased way, then subsequently used to update the stereotyping model. In this way, small quantities of evidence can rapidly result in polarized stereotypical opinions after the model is updated. In order to avoid this self-reinforcing feedback loop, we maintain unbiased opinions and stereotypical base rates separately, and only integrate them at the evaluation stage.

## 4. STEREOTYPE MODEL

A number of cognitive scientists have argued [McCauley 1994; Hilton and Von Hippel 1996] that prior probabilities (also referred to as base rates) in Bayesian inference are an appropriate way of describing the operation of stereotypes on an individual's beliefs. By representing stereotypes as base rates, their influence on a probability expectation should diminish as more direct evidence is observed and more reputational evidence is received.

Based on our notion of stereotypes from Section 2, we can consider a stereotype pragmatically as a rule, or set of rules, assigning some prior estimate of trustworthiness to individuals, based on the features they are observed to possess. Therefore, we can model the stereotypical evaluation procedure of an agent $x$, with respect to a particular task $\tau \in \mathcal{T}$ as a function $S_\tau^x : 2^{\mathcal{F}} \to \mathbb{R}$ mapping possible feature vectors of agents to initial stereotypical base rate estimates for those agents.

In this way, all of the stereotypes held by an individual are represented by one function. Since we have taken a probabilistic view of trust here, it should be possible to incorporate these estimates into any probabilistic trust model, and so the output of $S_\tau^x$ must be normalized. This allows us to evaluate a model of stereotypes in the context of a general trust evaluation mechanism. However, it is important to note that this need not be the case. In this case, the function's output would not represent a base rate, but an estimate appropriate for the trust metric being used, such as the most likely discrete trust value. Regardless of the underlying model, the key requirement for the stereotyping function is that the estimates it produces are compatible with the trust evaluation model being used.

The base rate parameter in SL then provides an appropriate way to incorporate the predictions of our stereotyping model back into the trust evaluation process. We model the effects of stereotypes in SL by using the model's predictions as the base rate. That is, for a given agent $y$, the base rate $a_{y:\tau}^x = S_x(F_y, \tau)$. For example, when no evidence has been received for a particular trustee, we have maximum ambiguity, that is, $\omega_{y:\tau}^x = (0, 0, 1, 0.5)$. In this case, $a_{y:\tau}^x$ alone determines the value of $P(\omega_{y:\tau}^x)$. However, as more evidence is accrued, the value of $u_{y:\tau}^x$ decreases, and so the effect of $a_{y:\tau}^x$ also decreases. This satisfies our fundamental condition that a stereotypical assumption must yield to concrete evidence as that evidence is acquired.

Note that our formulation of stereotypes is *decentralized*. It is not necessary for agents to agree on a common stereotype, as each agent maintains its own stereotypical model. It is possible, however, for groups of agents to share stereotypical beliefs about other groups. These *stereotypical biases* are discussed in Section 5.

### 4.1. Stereotypical Reputation

Constructing stereotypes still requires a significant number of interactions to be accumulated. New trustors can, however, make use of *stereotypical reputation* gathered

from experienced trustors who have already constructed stereotypes. When evaluating a given agent $y$, a trustor $x$ will perform a stereotype query when the following conditions hold:

(1) $x$ has no direct evidence from which to produce an evaluation for $y$;
(2) no reputational evidence about $y$ can be found from the recommenders visible to $x$ ($R_x$);
(3) $x$ cannot form a stereotypical evaluation for $y$, that is, when $x$ has not directly interacted with any agents before, or has not observed any of the features in $F_y$ before.

In this case, $x$ can ask visible reputation providers if they are able to provide stereotypical evaluations of $y$, in lieu of a concrete opinion about $y$. Once all stereotypical ratings have been received, $x$ computes the weighted mean of all ratings, with ratings weighted by the confidence value of each provider's stereotyping function.

$$SR^x_{y:\tau} = \frac{\sum_{z \in R_x} c_z a^z_{y:\tau}}{\sum_{z \in R_x} c_z} \tag{9}$$

Eq. (9) shows how stereotypical reputation $SR^x_{y:\tau}$ is calculated as the weighted mean of all returned stereotypical evaluations, with each evaluation weighted by the confidence its respective provider places in its stereotype model $c_z$. $a^z_{y:\tau}$ denotes the stereotypical evaluation produced by some agent $z$ about another agent $y$ performing task $\tau$. In our approach, this is given by the Root Mean Squared Error (RMSE) [Willmott et al. 1985] of the stereotype model of an agent $z$, $S_z$. This provides a measure of the model's accuracy as a function of the differences between the actual opinions and those predicted by the model, as shown in Eq. (10).

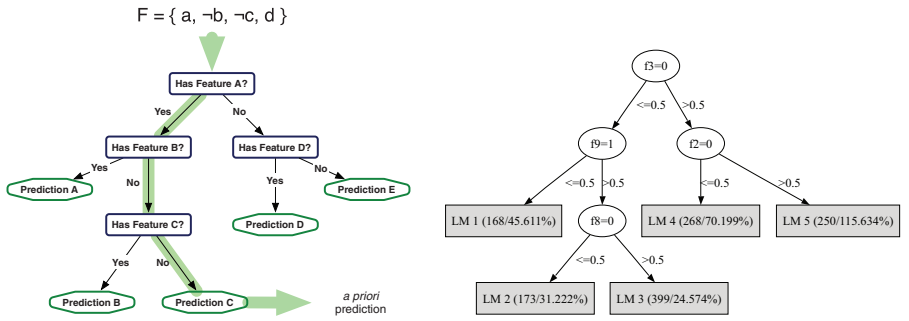
$$c_z = 1 - \sqrt{\frac{\sum_{\omega^z_{y:\tau} \in O_z} \left(P\left(\omega^z_{y:\tau}\right) - S_z(F_y, \tau)\right)^2}{|O_z|}} \tag{10}$$

In open MAS, it is possible that agents responding to queries may provide opinions and stereotypical evaluations which are biased in some way. For example, members of an organization may be inclined to provide positively biased stereotypical ratings of other agents from the same organization. We examine this problem in detail in Section 5. Before this, however, we describe how stereotypes are constructed from experiences with agents presenting similar features.

### 4.2. Learning Stereotypes

As we have mentioned, the ultimate goal of any stereotyping mechanism should be to learn a stereotyping function $S^a_\tau$ able to produce a priori assumptions based on the observable features of candidates. While the trust model we have described before is sufficient for the application of stereotypes and their integration with available direct and reputational evidence, we still require mechanisms for generalizing from known experiences to stereotypes.

Decision trees [Breiman 1984] provide an appropriate model for the behavior of stereotyping functions. By representing the stereotyping function in this way, we can make use of well-known techniques for inducing decision trees from labeled examples [Frank et al. 1998; Kalles and Morris 1996]. Furthermore, it is possible to




(a) example decision tree encoding a stereotyping function

(b) example M5 model tree

Fig. 3. Encoding stereotypes as decision trees.

encapsulate all of an agent's stereotypes about others regarding features in $\mathcal{F}$ in one concise structure. Each node of the tree represents a particular feature, and branches from nodes are followed depending on the perceived value of the feature represented by that node. Each leaf of the tree represents the stereotypical base rate (or a function producing a base rate) that will be applied to all classification examples reaching that leaf. Figure 3(a) shows an example of a simple decision tree being used to classify an agent with a visible feature vector $F = \{a, \neg b, \neg c, d\}$. The resulting path through the tree results in a predicted stereotypical evaluation for the agent, based on the feature vector $F$. When evaluating an agent $y$ for which we have no evidence, the stereotype tree can be used to obtain an estimated a priori trust value for $P(\omega^x_{y:\tau})$. With this value, we can create a new opinion about $y$, setting the predicted value as the base rate. This satisfies our requirements for the $S^x_\tau$ function.

Unfortunately, classical decision tree induction techniques are not suitable for problems where the class value to be predicted is real-valued. As we wish to predict a priori trust estimates, we must either perform some discretization of the range [0..1] to obtain discrete class labels, or use a decision tree induction technique which accommodates real-valued class labels. It is not immediately evident which of these approaches is most appropriate for our needs. Therefore, for comparison, we present and evaluate two different tree-based methods for inducing stereotypes. The first of these, which we term *two-phase learning*, involves labeling trustees from a finite set of discrete trust values. The second approach, *model tree learning*, involves learning a classifier capable of predicting exact trust values from a continuous range.

### 4.3. Two-Phase Learning

The former approach involves obtaining a set of labels by partitioning the space of opinions in some way, then tagging known agents with the label of the partition to which they belong. For example, a highly crude scheme may be to divide the partition space in two, labeling one half as "good" opinions, and the other as "bad". Alternatively, we can use a static partitioning scheme of some fixed granularity, such as that proposed by Jøsang and Pope [2005]. However, using a static scheme may result in the same behavioral patterns of agents being labeled differently, simply because opinions fall into different fixed partitions. A more flexible approach involves attempting to find statistically identifiable *clusters* of similar opinions using clustering techniques [MacKay 2003], and using the cluster membership of trustees as labels in a secondary *classification* stage.

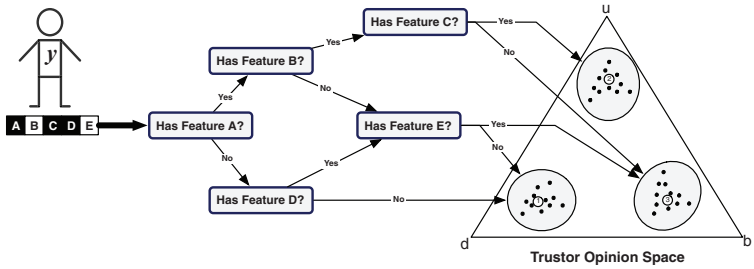



Fig. 4. Two-phase stereotyping mechanism.

The two-stage approach is illustrated in Figure 4 and proceeds as follows. In the first phase, a trustor obtains a set of discrete labels by identifying clusters[3] of trustees in its opinion base. The feature vector of each known trustee $y$ is then labeled according to its cluster membership $L_y$ to give a set of tuples $\langle F_y, L_y \rangle$. In the second phase, we construct a decision tree from these labeled examples produced by the first phase. For each opinion $\omega^x_{y:\tau} \in Ops_x$, unless $u^x_{y:\tau} = 1$ (totally ambiguous opinions add no knowledge to the model), we add the example $\langle F_y, L_y \rangle$ to the training set. The tree is then constructed. Since we would like to obtain base-rates from our model, it is necessary to convert class labels to some continuous value which is representative of their corresponding clusters. This is obtained for each label by taking the mean of all opinions (essentially the cluster centroid) within the cluster corresponding to that label. The rating function (Eq. (7)) is then used to convert these three-dimensional centroids to a single probability expectation value.

The two stages of the process are then as follows.

(1) A clustering stage produces a set of labels, representing observed behavioral classes.
(2) A classification stage attempts to learn correlations between the features of agents and their membership of the behavioral classes identified in the previous stage.

Since we would like to obtain base rates from our model, it is necessary to convert class labels to a continuous value which can be incorporated back into our opinion representation. This can be obtained for each label by taking the mean of all opinions within the partition corresponding to that particular label. The drawback to this approach is that the accuracy of the resulting base rate predictions may depend on the number of labels produced. For example, if we produce only two labels, the model's output may not be very informative, as only two point estimates will be available for the entire opinion space. However, by clustering opinions, we at least attempt to create a set of labels which are sensitive to the observed distribution of opinions.

### 4.4. Model Tree Learning

An alternative to two-phase learning is to use a decision tree induction technique capable of predicting numerical values. The M5 model tree learning algorithm [Quinlan 1992; Frank et al. 1998] is similar to other decision tree induction methods, in that it recursively constructs a tree for classification. However, while leaves of classical decision trees are class labels, the leaves of a model tree are linear regression models which are used to estimate the target value (in our case, a probability expectation value) as a function of the values of an agent's features. The approach differs from traditional tree induction methods in two key respects. Firstly, the "splitting criterion"

[3]In our evaluation of this method, we employed a k-means clustering algorithm [MacKay 2003] for this phase.

(by which the algorithm chooses which attributes to use as splitting nodes in the tree, or whether to split at all) used by most tree induction methods involves the maximization of information gain, by which attributes are selected that most rapidly reduce the "impurity" of the training data subsets created by splitting. M5, by contrast, selects attributes for splitting which minimize the standard deviation of the resulting subsets. Secondly, once a tree is constructed, linear models are computed for each leaf node of the tree, using standard regression techniques. A smoothing process then attempts to compensate for any sharp discontinuities between the resulting linear models.

This has the advantage of allowing a numerical estimate to be predicted directly, without necessitating the use of a clustering stage. Figure 3(b) shows an example model tree representing a learned stereotype, with agent features as nodes, feature values as paths, and linear models as leaves.

### 4.5. Unobserved Features

We previously mentioned that it may not be possible to observe the values of all features in a target agent's feature vector. In MAS, agents may attempt to hide their features, or may imperfectly perceive the feature vectors of others.

In such cases, where the observable feature information is limited, agents must make some assumptions about the most likely values of the unobservable features. It may be that the presence (or absence) of some features is highly correlated with others, and if so, reasonable assumptions about the most likely values of unobservable features may made through imputation [Quinlan 1986, 1993]. In decision trees, imputation is commonly performed by computing the most likely feature value among the training examples that reach the node representing the feature which is unobservable. In this way, the most likely value for the feature can be found, given the features which are visible.

## 5. STEREOTYPICAL BIASES

Until now we have used the term feature-behavior correlation to define the relationship between an agent's features and its trustworthiness in a given interaction. However, other kinds of feature-behavioral correlations may exist within a multiagent society, such as *biases* which affect the ways in which agents behave with and perceive their partners, depending on the features of both parties. We consider two main bias "types".

—*Perceptual bias*. Trustors may *perceive* a trustee's task outcomes more positively or negatively as a result of the trustee possessing (or lacking) certain features, that is, trustors use different subjective evaluation ($\varsigma_\tau^x$) functions depending on the trustee's features.
—*Behavioral bias*. Trustees with certain features *behave* more positively or negatively in interactions depending on the features of the trustor. For example, trustees sharing particular features may behave more reliably with one group of agents than with another.

These biases have the capacity to severely affect the performance of the reputational component of trust models. The presence of perceptual and behavioral biases means that trustors can no longer incorporate reputation from different sources easily. As the behaviors and perceptions of agents may depend on features of their partners, not all opinions will be appropriate for all trustors. By integrating reputational opinions without considering the possibility of social biases, agents may make erroneous decisions.

### 5.1. Reputation Filtering

In the presence of perceptual and behavioral biases within a society, we would expect that agents who naïvely aggregate biased opinions will form misleading trust evaluations, and hence perform more poorly. When gathering reputational evidence under bias, trustors should avoid integrating opinions that are deemed to be biased in some way. Unlike approaches which address the problem of deception [Sensoy et al. 2009; Jøsang and Ismail 2002], we do not assume that opinions that are divergent from the majority are necessarily inaccurate or deceptive. For example, if agents of a minority type are likely to share similar subjective evaluation functions (i.e., they share a perceptual bias), then those agents may choose to seek the opinions of feature-similar opinion providers who are more likely to provide opinions appropriate for them, even though these providers would be considered outliers by naïve reputation aggregation. This will also be the case when some trustees are behaviorally biased against the minority group. We take a simple approach, whereby opinion providers deemed to be uninformative due to social biases are omitted from the reputation aggregation process. Our intuition is that biased recommenders provide opinions that do not reflect the outcome a querying agent can expect from a trustee. If stereotypical biases are detected, agents will select the set of recommenders they perceive to be most appropriate, based on their own features and those of the recommenders.

We can therefore describe the reputational filtering process of an agent $x$ for a task $\tau$ as a function $\phi_\tau^x : 2^A \rightarrow 2^A$, which, given a subset of possible recommenders, returns a further subset of those recommenders appropriately filtered according to the perceived stereotypical biases.

We employ the two-stage learning approach outlined in Section 4.3. The main difference is that we are now attempting to find relationships between the features of *reputation providers* (as opposed to trustors) and the subjective opinions they have about trustees. The clustering stage is now responsible for identifying significant behavioral or perceptual variations among agents, as opposed to producing a sensible discretization scheme.

The general procedure is as follows. Firstly, a trustor $x$ queries the set of visible reputation providers $R_x$ to obtain a set of opinions $Ops_{y:\tau}^{R_x}$, and attempts to find clusters of opinions which may indicate the presence of behavioral or perceptual biases. For each recommender $z \in R_x$, we label that agent's feature vector, $F_z$, according to the cluster to which it belongs. We then use these labeled vectors to build a classifier from features to cluster membership. Finally, the trustor uses this tree to classify *itself* according to its own feature vector. The trustees ultimately selected for querying are those whose opinions fall within the same cluster as the trustor, according to the decision tree. As we will show, this approach can detect both behavioral and perceptual biases.

## 6. EVALUATION

In order to evaluate the effectiveness of our approach, we implemented the framework outlined before within a simulated multiagent system in which agents join, leave, interact, and share experiences over a number of interaction rounds. We investigate the following hypotheses.

—*Hypothesis* 1. If feature-behavioral correlations are present, then stereotyping agents will perform better than nonstereotyping agents.
—*Hypothesis* 2. The performance of stereotyping models will decrease as the strength of feature-behavioral correlations decreases.
—*Hypothesis* 3. If no feature-behavioral correlations exist, then stereotyping agents will perform no worse than nonstereotyping agents.

Table I. Trustee Profiles

| ID | Description | Mean | StDev | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ | $+noise$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $p_1$ | Usually good | 0.9 | 0.05 | x | | | | | x | ... |
| $p_2$ | Often good | 0.6 | 0.15 | | x | | x | | | ... |
| $p_3$ | Often bad | 0.4 | 0.15 | | | x | x | | | ... |
| $p_4$ | Usually bad | 0.3 | 0.05 | | x | x | | x | | ... |
| $p_5$ | Random | 0.5 | 1.0 | | x | x | | | x | ... |

—*Hypothesis* 4. If either perceptual or behavioral biases are present, then reputation filtering agents will perform better than nonreputation filtering agents.

In the following section, we will outline our experimental framework and present our results with respect to these hypotheses.

### 6.1. Stereotyping Experiments

In our experiments, we create a fixed number of agents, and assign to each agent the role of either trustee or trustor. While it is not necessary for roles to be fixed, this allows us to clearly assess the impact of the different trust models on the quality of the trustors' evaluations. Specifically, 500 agents are created to play the role of trustees, and 40 agents to play the role of trustors. We create 20 ad hoc groups within the society, each comprising 10 agents. The mixture of trustors and trustees in each group is randomly determined. Therefore, some groups may have more trustors than trustees, and vice versa. Groups comprising agents of only one role will not engage in any interactions. Each ad hoc group exists for 5 interaction steps, after which it is disbanded, and a new group created in its place. In each interaction step, each trustor agent interacts with a trustee with an interaction probability $P(interact) = 0.8$. Also, we control the basic rate of dynamicity in the society with a join/leave probability parameter $P(jl) = 0.01$, which determines the probability with which, in each interaction step, a trustee will leave the society, to be immediately replaced by a new trustee from the same profile. Each experiment lasts for 400 interaction steps.

Trustees are drawn from a number of hidden profiles which determine their behavioral characteristics. 100 trustees from each profile were created. By creating an even distribution of agent profiles, we aimed to minimize any effect caused by trustors being more or less likely to encounter trustees from one profile than another. Also, due to the level of dynamism in the simulation, some agents may find themselves assigned to ad hoc groups comprising only good or bad partners. As a result, the performance of individual agents may be affected by chance as well as the performance of their respective trust models. For this reason, we use the global average interaction outcome at each time step as a performance metric.

Each profile specifies the mean and standard deviation parameters of a Gaussian distribution from which simulated interaction outcomes will be drawn, so that $O_\tau = \{o|0 \leq o \leq 1\}$. We define the set of possible features $\mathcal{F} = \{f_1 \ldots f_{18}\}$, and each profile specifies the values of the first 6 features (termed *diagnostic* features) from $\mathcal{F}$ which are shared by all agents of that profile. In this way, we define the feature-behavior relationships we wish our agents to identify. All features are represented as binary variables, each signifying the presence or absence of a given feature. The remaining 12 features in $\mathcal{F}$ represent *noise* features, which are randomly assigned and do not correlate with profile membership. We assume here that all agents share the same subjective evaluation function $\varsigma_\tau^x$ for all tasks, which evaluates any outcome $o_\tau \geq 0.5$ as a positive outcome, and negative otherwise.

The test profiles used in our experiments are given in Table I. The profile $p_1$ represents a reliable class of agents, while $p_4$ represents agents who will usually perform

poorly. Profiles $p_2$ and $p_3$ represent unreliable agents who may perform well or poorly, and $p_5$ represents agents with uniform performance distributions. Agents of type $p_5$ add noise, as their behavior is evenly distributed on either side of the threshold value of 0.5.

In evaluating the stereotypical reputation function, we employ three experimental conditions.

(1) *Global interaction and reputation*. Trustors can select any partner from the *global* society to interact with, and can query the global society for reputational opinions.
(2) *Ad hoc interaction, global reputation*. Trustors can only interact within their ad hoc groups, but can query the global society for reputational opinions.
(3) *Ad hoc interaction and reputation*. Trustors can only interact and communicate within their ad hoc groups.

In each condition, we compare the performance of the nonstereotyping trust evaluation model with the same model employing a stereotyping approach. Both the two-phase and M5 tree approaches are compared.

### 6.2. Stereotyping Results

All graphs (except for Figure 7(b)) plot the mean objective interaction outcome, or social welfare, of the trustor population as a whole at each interaction. All results presented in this section have been found statistically significant by $t$-test with $p < 0.05$[4].

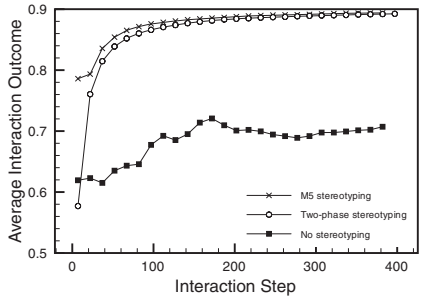
*6.2.1. Hypothesis 1.* Figures 5(a), 5(b), and 5(c) show the performance of our approach in conditions 1, 2, and 3, respectively. Both two-phase and M5 stereotyping models outperformed the nonstereotyping model after the first learning interval, with M5 achieving the best performance. This means that stereotyping agents are (in general) able to make better trust evaluations than their nonstereotyping counterparts. Condition 1 represents the least dynamic environment; only the $P(jl)$ parameter poses a challenge. Stereotyping agents are able to attain better performance here by reusing their experiences (through generalization) with known agents even after these agents have left the system. The stereotyping models demonstrate a significant benefit in both conditions 2 and 3, but the rate of increase of this benefit is reduced in condition 3, as agents have access to fewer sources of reputational opinions.

Figure 6(a) shows the performance of our approach when the $P(jl)$ parameter is increased to 0.5. This means that, in each interaction step, each trustee may be replaced with a probability of 0.5. Our results show that the stereotyping agents continue to perform well by making use of generalized trust evaluations. On the other hand, nonstereotyping agents rarely have the opportunity to use direct or reputational experiences.

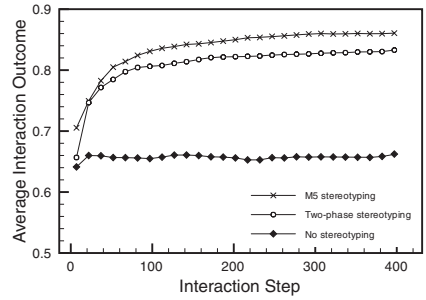
*6.2.2. Hypothesis 2.* In evaluating this hypothesis, we assign profile features probabilistically. Agents possess their diagnostic features with a probability of $p$, and other features with a probability of $1 - p$. This represents a weakening of the feature-behavioral correlation. The M5 model was evaluated with various values of $p$. As expected, Figure 5(d) shows that the benefit provided by the stereotyping models is diminished when features are not always predictive. This confirms our hypothesis that the effects of the stereotyping model will diminish as features become less predictive of a trustee's profile.

*6.2.3. Hypothesis 3.* By setting all features to be "noise" features (i.e., randomly assigned) we remove feature-behavioral correlations entirely. This is equivalent to the worst case in Figure 5(d), where $p = 0.5$. Figure 6(b) shows that neither stereotyping

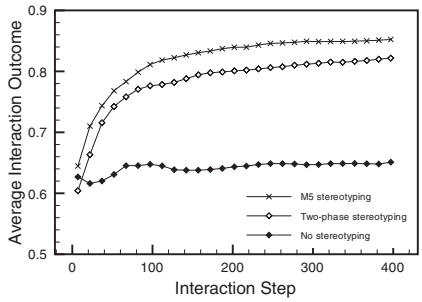
[4]Tests of statistical significance were carried out using the GNU R statistical computing environment [Hornik 2010].



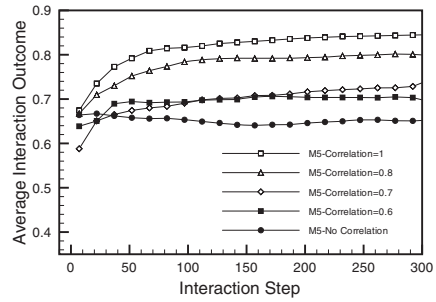

(a) global interaction and reputation




(b) ad hoc group interaction and global reputation

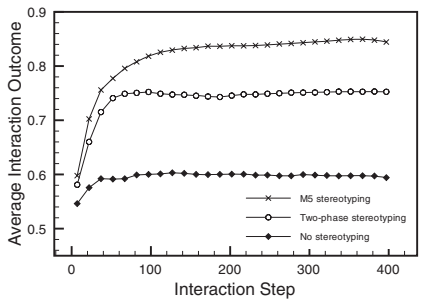



(c) ad hoc group interaction and reputation

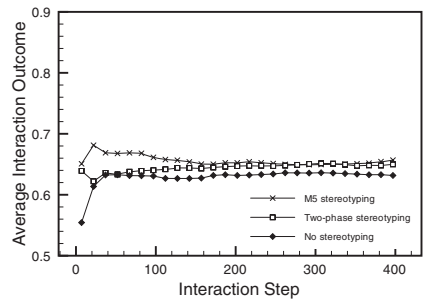



(d) unreliable feature-behavior correlations

Fig. 5. Stereotyping experimental results.




(a) high dynamicity, $P(jl) = 0.5$



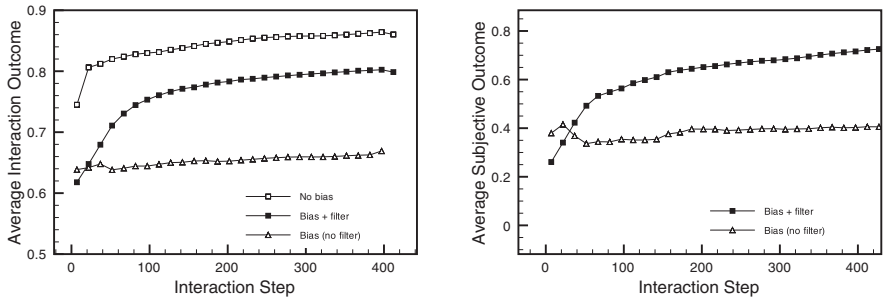

(b) no diagnostic features

Fig. 6. Stereotyping experimental results.

approach performs significantly better (or importantly, worse) than the nonstereotyping model when these correlations are not present. This confirms our hypothesis that using a stereotyping model will not lead to a degradation of performance if the assumption that feature-behavioral correlations exist does not hold.

Table II. Test Behavioral Biases

| Profile | Rule |
|---|---|
| $p_1$ | **IF** $f_2 \wedge f_4$ **THEN** $\langle \bar{m} = 0.3, \sigma = 0.05 \rangle$ **ELSE** $\langle \bar{m} = 0.8, \sigma = 0.05 \rangle$ |
| $p_2$ | **IF** $f_1 \wedge f_6$ **THEN** $\langle \bar{m} = 0.3, \sigma = 0.05 \rangle$ **ELSE** $\langle \bar{m} = 0.8, \sigma = 0.05 \rangle$ |




(a) behavioral biases and filtering (b) perceptual biases and filtering

Fig. 7. Bias filtering experimental results.

### 6.3. Bias Experiments

In evaluating our reputation filtering model, we ascribe behavioral and perceptual biases to profiles, in addition to the Gaussian parameters as before. We make use of two profiles, $p1$ and $p2$, where each profile is negatively biased (both behaviorally and perceptually) towards the other, and positively biased towards agents with similar features.

*6.3.1. Behavioral Biases.* Since we no longer assume trustees behave in a uniform way with all trustors, we model biased behavior of trustees towards trustors as rules that specify Gaussian parameters (mean $\bar{m}$ and standard deviation $\sigma$) to be used depending on the observed features of the trustor. Table II details the test biases used in our experiments. These biases mean that trustees will behave differently, depending on the features of trustors.

*6.3.2. Perceptual Biases.* Since trustors may be stereotypically biased in the way they *perceive* the performance of trustees, we model biased perception of trustors as rules that specify a different threshold value for the $\varsigma_\tau^x$ function to be used, depending on the features of the trustee. In our evaluation, we set the threshold to 0.6 when trustors evaluate trustees of the same profile, and 0.3 otherwise. This has the effect of making trustors' satisfaction criteria partially dependant on features of the trustee. In order to isolate the effect of perceptual biases, we set the $\bar{m}$ and $\sigma$ behavioral parameters of all trustees to 0.5 and 0.05, respectively.

### 6.4. Bias Results

*6.4.1. Hypothesis 4.* Figures 7(a) and 7(b) show the results of our bias experiments. Behavioral and perceptual biases were evaluated separately. Both graphs show the performance of the standard (nonstereotyping) trust model in the ideal case when no bias is present. However, when biases are present, the naïve model performs dramatically worse, only achieving an outcome slightly higher than what would be expected by chance. By using the reputation filtering mechanism, agents are able to mitigate some of the negative effects of behavioral and perceptual biases, and perform significantly better than naïve reputation aggregation.

Note that Figure 7(b) plots the average subjective outcome obtained trustors, as opposed to the actual observed task outcome. This is because trustors' possess different subjective evaluation functions, and so the effectiveness of the trust model can no longer be measured in terms of the objective outcome.

## 7. DISCUSSION

Our results show that stereotyping can offer a significant improvement under the conditions outlined in Section 1. Also, of the two approaches we presented, M5 seems to perform consistently better than the two-phase approach. This is, we believe, due to the precision lost when using cluster centroids as base rates. M5, on the other hand, attempts to construct linear models which are able to more accurately describe the relationships between features and behavior.

While we have referred to a number of trust evaluation models in this article, it is worth highlighting here some related approaches which attempt to address the issues of specific interest. The FIRE [Huynh et al. 2006] system employs *role-based trust* to explicitly capture relationships between agents in certain roles. Tailored rules specify an initial degree of trust that will be conferred on partners for whom the rules match. This means that a degree of trust may be present even when no evidence is available. In contrast with our approach, which learns stereotyping rules from observations, FIRE rules are explicitly specified for a domain at design time. Similarly, *system* trust in the REGRET [Sabater 2003] framework attempts to incorporate information about social categories, but again assumes these are provided in the form of rules by the system designer. The stereotyping approach presented here could be adapted to complement these existing mechanisms. However, with nonprobabilistic trust models, it is necessary to provide some scheme which allows stereotypes to be integrated alongside the trust dimensions considered by the mechanism in question, for example, through the use of weights.

While several authors have attempted to address the general issue of deceptive reputation providers [Sensoy et al. 2009; Yu and Singh 2003; Teacy et al. 2006], these approaches involve either learning about the trustworthiness of reputation providers (in a similar manner to learning about trustees), or treating statistically "outlying" opinions as untrustworthy. However, in dynamic societies, the turnover among reputation providers may be high, and so it may not be feasible to build models of individual provider trustworthiness. Similarly, simply filtering outlying or minority opinions may not always be appropriate for agents who are biased, or subject to the biases of others. While we do not attempt to provide a general solution to this problem here, our approach does not require the trustworthiness of reputation providers to be learned, nor do we assume that outlying opinions are necessarily untrustworthy.

One drawback with the filtering mechanism we have proposed is that the procedure of selecting appropriate reputation providers must be performed individually for each candidate-task combination. A more effective approach may be to extend the mechanism to permit further generalization to a notion of *interaction* stereotypes, representing patterns such as "agents with features $f_3$, $f_7$, and $f_8$ behave positively in task $\tau$ toward agents with features $f_1$ and $f_7$". This allows the mechanism to learn about biases in general, and extends the applicability of our approach to other tasks, such as trustworthy team formation. Knowledge of biases is desirable when it is necessary to entrust a complex task to a team of diverse agents, and when the success of the task depends on the successful collaboration of the constituent trustees. Future work will investigate the applicability of learned biases to the problem of trustworthy team formation.

Another key future direction involves exploiting ontological relationships between the features agents possess. In this work, we have assumed that all features are

independent of each other. In reality, however, there may be a large number of features for which hierarchical (or other) relationships exist, and these relationships could be useful when forming stereotypes. For example, the features "cardiologist", "general practitioner", and "surgeon" could all be considered subtypes of a more general feature "doctor". If feature-behavior correlations exist between these higher-level features (such as "all doctors are trustworthy at administering first-aid"), then it may be beneficial to exploit these, rather than constructing separate models for each of the more specific features. While agents may not be able to observe these high-level features directly, they can learn generalizations when they have access to ontological knowledge about the relationships between features. Future work will investigate ways in which knowledge of higher-level features could be used to produce more effective stereotypes.

Finally, it is worth noting that interesting cases may arise when an individual possesses features which match more than one stereotype. For example, a society may share a stereotype that football players, being naturally more concerned with physical rather than intellectual activities, cannot be expected to be competent at authoring books. On the other hand, successful authors, who have published a number of books, would be stereotyped as competent authors. How should a stereotyping model assess the competence of an agent (as an author) who is both an accomplished footballer and has successfully published a number of books?

In our current model, the final classification depends on the attributes chosen to be most predictive, given the available observations. In these conflicting circumstances, given that no football-playing authors have been previously observed, the stronger stereotype, with respect to a given task, will determine the final classification. For example, given that many books are written by successful authors, features pertaining to authors (such as their publisher, number of published books, etc.) will likely be more powerful predictors of competence than those pertaining to sporting ability. However, if sufficient experiences with football-playing authors can be obtained, a new stereotype can be formed to describe this class. It may be, for example, that football-playing authors generally write poor-quality books. On the other hand, it may be that they tend to write excellent books, for which their footballing skills suffer. Therefore, while our stereotyping model may deal crudely with such conflicting cases to begin with, new and informative stereotypes can be constructed to address them as more experiences are obtained.

## 8. CONCLUSIONS

In highly dynamic human societies, stereotyping-like processes are crucial to provide confidence to take initial risks which lay the foundations for the formation of trust. In their seminal work on swift trust, Meyerson et al. concluded that, in ad hoc teams, "people have to wade in on trust rather than wait while experience gradually shows who can be trusted and with what: Trust must be conferred presumptively or ex ante" [Meyerson et al. 1996]. We have shown that, much like their human analogs, highly dynamic virtual societies present a serious barrier to the formation of trust. The approach presented here can facilitate better initial trust evaluations when feature-behavioral correlations are present, by allowing agents to generalize from trust in individuals to trust in observable features. Where hidden feature-behavior correlations exist in the trustee population, our model has been shown robust when both interaction and reputation gathering were constrained to within ad hoc groups. Our model also performs well when the probability of agents leaving, joining, or changing identity is high.

We have also shown how stereotyping can help agents select appropriate reputation providers when stereotypical biases exist in the society. When these factors are not present, the stereotyping approach incurs no loss of performance. We have demonstrated how a stereotyping approach can be used together with a relatively

straightforward probabilistic trust model in order to significantly improve performance. However, the significance of our model lies in its immediate applicability to the current family of probabilistic trust models for MAS. As our stereotyping model learns from real-valued trust ratings, it is directly compatible with any model which uses numerical measures of trust.

## ACKNOWLEDGMENTS

We would like to extend our thanks to the reviewers, for their detailed and insightful comments towards improving this article.

## REFERENCES

Breiman, L. 1984. *Classification and Regression Trees*. Chapman & Hall.

Burnett, C., Norman, T. J., and Sycara, K. 2010. Bootstrapping trust evaluations through stereotypes. In *Proceedings of 9th International Conference on Autonomous Agents and Multiagent Systems*. 241–248.

Carver, L. and Turoff, M. 2007. Human-Computer interaction: The human and computer as a team in emergency management information systems. *Comm. ACM 50,* 3, 33–38.

Castelfranchi, C. and Falcone, R. 1998. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings of the 3rd International Conference on Multi Agent Systems*. 72–79.

Chen, M. and Bargh, J. 1997. Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *J. Experi. Social Psychol. 33*, 541–560.

Eschenauer, L., Gligor, V., and Baras, J. 2003. On trust establishment in mobile ad-hoc networks. In *Security Protocols*, Springer, 47–66.

Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. 1998. Using model trees for classification. *Mach. Learn. 32,* 1, 63–76.

Gambetta, D. 1990. *Trust: Making and Breaking Cooperative Relations*. Blackwell.

Hang, C. W., Wang, Y., and Singh, M. 2009. Operators for propagating trust and their evaluation in social networks. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems.* Vol. 2. 1025–1032.

Hilton, J. and Von Hippel, W. 1996. Stereotypes. *Annual Rev. Psychol. 47,* 1, 237–271.

Hornik, K. 2010. The R FAQ. http://CRAN.R-project.org/doc/FAQ/R-FAQ.html.

Huynh, T. D., Jennings, N. R., and Shadbolt, N. 2006. An integrated trust and reputation model for open multi-agent systems. *Auton. Agents Multi-Agent Syst. 13,* 2, 119–154.

Jarvenpaa, S. and Leidner, D. 1999. Communication and trust in global virtual teams. *Organiz. Sci. 10,* 6, 791–815.

Jøsang, A., Hayward, R., and Pope, S. 2006. Trust network analysis with subjective logic. In *Proceedings of the 29th Australasian Computer Science Conference*. Vol. 48, Australian Computer Society, 85–94.

Jøsang, A. and Ismail, R. 2002. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*.

Jøsang, A., Ismail, R., and Boyd, C. 2007. A survey of trust and reputation systems for online service provision. *Decis. Support Syst. 43,* 2, 618–644.

Jøsang, A. and Pope, S. 2005. Semantic constraints for trust transitivity. In *Proceedings of the 2nd Asia-Pacific Conference on Conceptual Modelling*. Vol. 43. 59–68.

Kalles, D. and Morris, T. 1996. Efficient incremental induction of decision trees. *Mach. Learn. 24,* 3, 231–242.

Lorini, E. and Demolombe, R. 2008. From binary trust to graded trust in information sources: a logical perspective. In *Trust in Agent Societies*, Springer, 205–225.

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

McCauley, C. 1994. Stereotypes as base rate predictions: commentary on Koehler on base-rate. *Psycol. 5*, 1055–1143.

Meyerson, D., Weick, K., and Kramer, R. 1996. Swift trust and temporary groups. In *Trust in Organizations: Frontiers of Theory and Research*, R. Kramer and T. Tyler, Eds., Sage Publications, 415–445.

Militello, L. G., Patterson, E. S., Bowman, L., and Wears, R. 2007. Information flow during crisis management: Challenges to coordination in the emergency operations center. *Cogn. Technol. Work 9,* 1, 25–31.

Quinlan, J. 1986. Induction of decision trees. *Mach. learn. 1,* 1, 81–106.

Quinlan, J. 1992. Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*. 343–348.

QUINLAN, J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, CA.

SABATER, J. 2003. Trust and reputation for agent societies. Ph.D. thesis, l'Institut d'Investigació en Intel·ligència Artificial.

SENSOY, M., ZHANG, J., YOLUM, P., AND COHEN, R. 2009. Poyraz: Context-Aware service selection under deception. *Comput. Intell. 25,* 4, 335–364.

TEACY, W., PATEL, J., JENNINGS, N. R., AND LUCK, M. 2006. Travos: Trust and reputation in the context of inaccurate information sources. *Auton. Agents Multi-Agent Syst. 12,* 2, 183–198.

WANG, Y. AND SINGH, M. P. 2007. Formal trust model for multiagent systems. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. 1551–1556.

WILLMOTT, C., ACKLESON, S., DAVIS, R., FEDDEMA, J., KLINK, K., LEGATES, D., O'DONNELL, J., AND ROWE, C. 1985. Statistics for the evaluation and comparison of models. *J. Geophys. Res. 90,* C5, 8995–9005.

YU, B. AND SINGH, M. 2003. Detecting deception in reputation management. In *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems*. ACM, 73–80.

Received August 2010; revised December 2010; accepted February 2011