Supporting Trust Assessment and Decision Making in Coalitions

Chris Burnett, *University of Aberdeen*

Timothy J. Norman, *University of Aberdeen*

Katia Sycara, *Carnegie Mellon University*

Nir Oren, *University of Aberdeen*

Although multiorganizational coalitions offer diverse capabilities, assets, and information sources, successfully completing dynamic operations can compromise coalition partners' trust. Mechanisms for assessing trust can help, along with controls to help decide the best course of action.

Modern coalition operations require multiple organizations to act in a coordinated manner to achieve some high-level goal. Such operations therefore require trust at multiple levels. In military coalitions, for example, one unit trusts that others will be in position at the right time to provide support. Within modern coalitions—often composed of multiple nations and nongovernmental organizations—units rely on each other to provide information relevant to their shared objectives, but that information might be summarized or redacted between partners who don't trust one another. Trust, therefore, plays a critical role in deciding whether to share information and how to go about acquiring it.

Consider the following scenario: Two nations, A and B, are part of a coalition attempting to stabilize a war-torn country. Other coalition members include the local police, army, and various agencies of both nations. International aid agencies, journalists, and local informants might also interact with the coalition. Although it's in the interest of all such actors to share information and information-acquisition assets (such as sensors or human sources), the dynamic nature of the trust relationships between them can impede effective information sharing and acquisition.

Let's say, for example, that nation A wants to obtain aerial imagery of a particular area before taking some critical action. Nation A owns limited sensing assets (say, an unmanned aerial vehicle) that, given time, can be deployed to the area. However, other coalition partners have sensing assets already in the area that can provide the aerial imagery, and might even be in a position to share relevant information immediately. Now, Nation A must consider carefully whether to deploy its own resources, which entails a time cost, or delegate the sensing action to another coalition partner, exposing A to the risk of obtaining inaccurate information.

Because a coalition's success can hinge on its ability to effectively delegate tasks and

1541-1672/14/$31.00 © 2014 IEEE
Published by the IEEE Computer Society

share information and assets, there's a need for trust-aware decision-support tools that explicitly factor trust into the decision-making process. Here, we discuss work toward a probabilistic trust-assessment mechanism that can be used to determine the trustworthiness of an information source based on its historical accuracy, as well as that of other *similar* sources. We then discuss a decision-theoretic mechanism for making decisions about whether to trust an information source based on trust assessments, and, if so, how to tailor the interaction so as to minimize the perceived risk.

Trust Assessment and Decision Making

In general, any comprehensive trust-support mechanism will comprise three phases (see Figure 1):

- trust assessment, in which trust levels are evaluated using available evidence;
- decision making, in which trust levels are used to make decisions within a particular context, taking advantage of available sanctions and incentives; and
- trust update, in which post-hoc observations are integrated to inform future assessments.

In the context of information sharing within a coalition, these phases correspond to:

- evaluating available information sources in a way that allows them to be compared;
- making decisions about how to acquire information, perhaps using additional information sources and incentives to increase confidence in the information's reliability; and
- updating trust assessments of agents on the basis of the provided information's perceived trustworthiness, through, for example, an *a posteriori* comparison with observations of the ground truth.

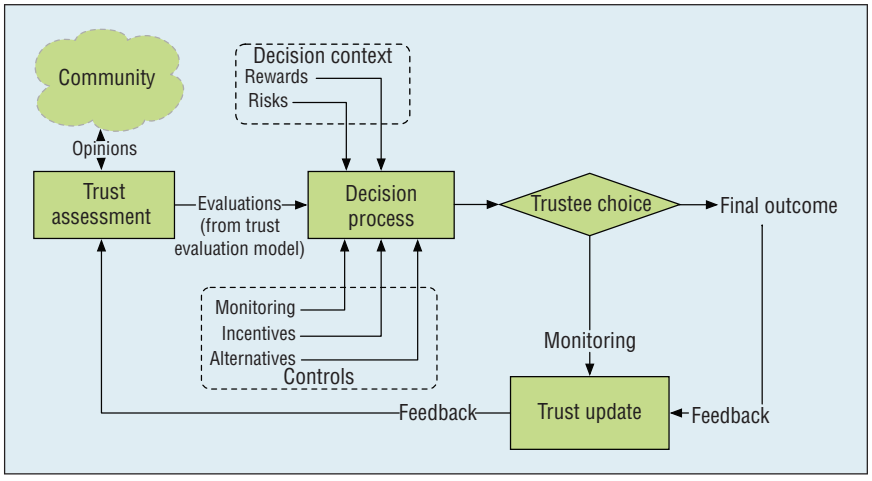

Figure 1. The trust-assessment and decision-making architecture.

For more information on aspects of the trust-assessment process, see the related sidebar.

Trust Assessment and Update

A key trust-assessment challenge in large, complex, and highly dynamic coalitions is that members frequently join and leave. Although a coalition's organizational structure might be relatively static, its members might be drawn from diverse pools of people who lack prior experience of working together and might come and go frequently. As a result, there's often insufficient direct or reputational evidence available about a particular individual to make a confident trust assessment. Without some trust assessment, agents might be unwilling to interact with the target; this prevents any evidence of the target's trustworthiness from being gathered and stalls the bootstrapping of trust. To address this problem, we've developed mechanisms that enable trust models to generalize from the available direct and reputational evidence about known agents to create *stereotypical* assessments about unknown agents.[1]

The concept of forming stereotypes is intuitive, inspired by human trust-assessment models such as Swift Trust,[2] which describe how trust is formed in a relationship's early stages. Generally, a stereotype is a set of rules that, given a set of agent *features*, returns an estimated trust assessment. Examples of such features in a coalition setting include nationality, organizational memberships, location, and relationships with other coalition partners. To construct these rules, an agent annotates its existing trust assessments of other agents (in probabilistic approaches, this is typically a single number), with their observable features. The key challenge then lies in identifying which features are good predictors of an agent's trust assessments.

Machine learning techniques provide one way to do this. Our approach employs decision-tree induction methods[3] to induce decision trees (see Figure 2). These trees can be used to visually map a set of features to an a priori trust evaluation for an agent possessing those features. This process doesn't require any direct or second-hand evidence with the target agent; you simply need access to some of its observable features, and a body of evidence with other agents that share some predictive features with the target.

In our approach, we use an M5 model tree induction algorithm[3] to learn stereotypes. Model trees are

Aspects of the Trust-Assessment Process

Computational trust models have been the focus of extensive research in recent years. Rather than provide a survey here (a good overview of this field can be found elsewhere[1,2]), we outline fundamental concepts common to any comprehensive trust-assessment system.

Paradigms

Two popular paradigms have emerged for the development of computational trust models: *cognitive* and *probabilistic* paradigms. Christiano Castelfranchi and Rino Falcone[3] describe a cognitive approach in which trust is viewed as the different beliefs about the trusted agent's internal mental state, such as its innate competence and intentional willingness (that is, its intentions and commitment) in relation to achieving some goal. In contrast, probabilistic approaches assess agents based on observations of their external behavior. Such approaches form the foundation of many existing state-of-the-art trust-assessment mechanisms, including our own.

Evidence

Regardless of the underlying paradigm, any trust model will be concerned with the aggregation of *evidence* from which assessments can be made. We identify three general classes of evidence: direct, second-hand, and stereotypical. *Direct evidence* concerns an agent's own past experiences, whereas *second-hand evidence* refers to opinions received from others. When direct or second-hand evidence is unavailable, observed correlations between agents' visible features and their behaviors constitute a form of *stereotypical evidence*. Stereotypes are discussed in more detail below, as they're a key contributor to our approach.

Aggregation

Using second-hand evidence introduc es a new problem: how to find trustworthy opinions and aggregate them together with direct evidence to form a single, comparable trust metric. Aggregation techniques can be categorized as either exogenous or endogenous.[1] *Exogenous* approaches build trust models of opinion providers themselves, based on the accuracy of past opinions, and use this to weight opinions subsequently received. *Endogenous* approaches use statistical properties of the set of opinions received from numerous providers to determine which of them are trustworthy and which aren't. For example, agents with extreme, outlying opinions might be considered inaccurate by an endogenous approach.

Biases

The behavior of a coalition's partners can be affected by various biases. Partners might behave in a more (or less) trustworthy manner when interacting with certain others; we refer to this as *behavioral bias.* On the other hand, partners might perceive the behaviors of certain others more (or less) favorably; we refer to this as *perceptual bias.* These biases can also manifest stereotypically; for example, an agent might behave more favorably when interacting with others who share some of its features. Perceptual and behavioral biases present a problem for second-hand evidence aggregation, as the naïve aggregation of biased opinions can result in skewed trust assessments. Trust-alignment techniques[4] attempt to address these issues by enabling agents to identify differences between their trust models and calculate alignments—that is, mappings between different trust models—to negate bias.

References

1. A. Jøsang, R. Ismail, and C. Boyd, "A Survey of Trust and Reputation Systems for Online Service Provision," *Decision Support Systems*, vol. 43, no. 2, 2007, pp. 618–644.
2. J. Sabater and C. Sierra, "Review on Computational Trust and Reputation Models," *Artificial Intelligence Rev.*, vol. 24, no. 1, 2005, pp. 33–60.
3. C. Castelfranchi and R. Falcone, "Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification," *Proc. 3rd Int'l Conf. Multi-Agent Systems*, 1998, pp. 72–79.
4. A. Koster, M. Schorlemmer, and J. Sabater-Mir, "Engineering Trust Alignment: Theory, Method and Experimentation," *Int'l J. Human-Computer Studies*, vol. 70, no. 6, 2012, pp. 450–473.

decision trees with linear regression models at leaf nodes that can be used to predict continuous values (in our case, a priori trust estimates). Although we could use other machine learning techniques, decision trees have the advantage of being visually inspected, which is crucial for informing the human decision-making processes. These stereotypical predictions can assist in the "bootstrapping," or initialization, of trust relationships by allowing agents to form tentative assessments that encourage initial interactions. Stereotypes can also be shared with other agents in the society, leading to a form of *stereotypical reputation*. Interestingly, this means that the behavior of one agent with a given feature might have an effect on the trust subsequently afforded by other agents to those sharing that feature.

Trust Decision Making

Once a given situation's set of trust assessments has been computed, it's then necessary to evaluate the alternatives available, considering the risks and rewards associated with each. Coalition partners have a degree of control over the levels of service they provide. For example, some partners might wish to share information, trusting their counterparts to use the information effectively to benefit the coalition. However, some partners might not trust others to keep certain sensitive aspects confidential, and so might choose to deviate from prior agreements about information quality by providing less reliable data or by obfuscating their messages. Two different trust issues, or *contexts*, are involved here: a partner might be trusted to put confidential information to good use, but this doesn't

necessarily imply trust that the partner will refrain from sharing that information with others. Here, we assume that these two trust contexts are assessed separately.

Economic models of delegation, such as the Principal-Agent Theory, provide a good starting point for modeling these situations.[4] Our mechanism builds on aspects of Principal-Agent Theory to provide trustors with several controls—such as sanctions, rewards, and monitoring strategies—that let them influence the trustee's choice of action and learn about the trustee's behavior given a particular choice.

We consider trustees who have various actions available to them (for example, providing high or low service), each with a different probability of yielding a particular outcome o (for example, *success* or *failure* in locating a particular asset). In coalition settings, actions more likely to yield positive outcomes for the trustor will often be more expensive or risky for the trustee, resulting in a moral hazard. For example, providing high-resolution sensor data of resource positions might expose the provider to risk if that data were to be passed on to unauthorized recipients. These choices are private by default. Now, for each trustee, we must build and maintain a set of *conditional* trust models that estimate trust given a particular action choice. This requires observations of agents' action choices, which can be obtained only through costly *monitoring*.

Because trustors might need to motivate trustees to choose desirable actions, delegation in such a context requires trustors to compute a *contract* function specifying a positive or negative payoff to be transferred between the trustor (that is, the information consumer) and the trustee (the information producer) given some observed task outcome o (such as the provided information's *post-hoc* quality in relation to prior expectations) or an observed choice of trustee actions (visible through monitoring).

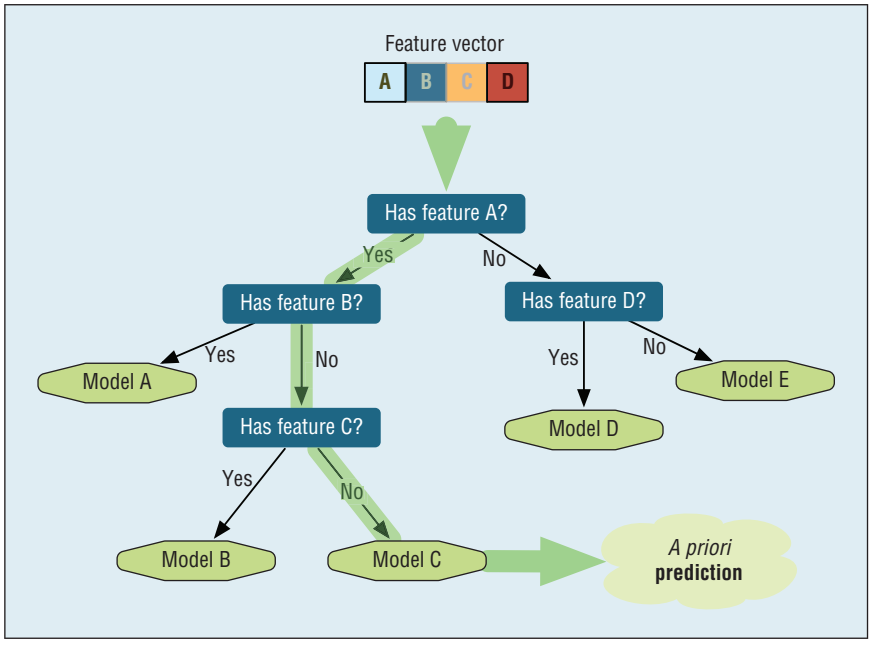

Figure 2. A stereotype tree.

Our mechanism considers three types of contract: incentive, monitoring, and reputational incentive (RI). In an *incentive contract*, the trustor creates a contract specifying the trustee's compensation dependent on the *outcome*, which is a function of the trustee's choice. In a coalition setting, these incentives could include monetary transfers, the granting/revoking of access to coalition resources, or the adjustment of existing information sharing agreements, all of which can affect the trustee's welfare.

In a *monitoring contract*, the trustor expends additional effort or utility to observe the trustee's behavioral choices. This might involve acquiring corroborating evidence from trusted partners to identify any deliberate obfuscation. Monitoring changes the contract's nature; payoffs might be contingent on the trustee's *observed action choice*. So, if the quality of received information deviates significantly from that expected—*and* from that received by other (trusted) corroborators—it might indicate that the deviation was the result of an intentional choice on the provider's part. However, monitoring actions aren't always available, and generally incur a significant cost on the trustor's part.

In an *RI contract*, the added incentive is the potential change in reputation that a trustee will experience as a result of feedback being communicated to the society, with the associated utility gain (or loss). Robert Axelrod and W.D. Hamilton called this the "shadow of the future" effect.[5] This contract requires a degree of competition between providers. Informally, RI is computed by estimating the change in expected utility a provider will obtain in future interactions, as a result of the consumer's feedback making the provider more (or less) likely to be selected for interaction. In a coalition, this is particularly salient, as a low reputation could mean expulsion from the coalition and the loss of associated benefits.

We formulated the computation of these contracts as a convex optimization problem. In our model, we used the Convex Optimization in Python[6] package to find optimal contracts efficiently.

For each potential candidate, three contracts (incentive, monitoring, and RI) must be computed. The problem grows with the number of candidates that must be evaluated, the number of actions available to them, and the number of trust contexts or tasks involved. Once the trustor has identified the candidate and contract that provides the highest expected return, interaction takes place. More formal details regarding these agreements can be found elsewhere.[7]

Returning to our introductory example, suppose that nation A wants to obtain some information from a local informant source. The coalition requires information about the location of a planned meeting between insurgent leaders. The informant indicates that he is capable of obtaining this information, but can privately choose to provide either high-quality, accurate information, or low-quality, less accurate information. For example, the informant can replace exact pieces of information (such as specific locations) with "fuzzy" ones (say, regions). Although the coalition would prefer high-quality information, providing it could put the informant at risk of being uncovered. Given this, the coalition can respond by using incentives to influence this choice, for example, or by contacting other information sources for corroborating evidence.

Evaluation

Our system implements the trust-assessment and decision-making process shown in Figure 1, and was evaluated in a multiagent simulation. We employ Subjective Logic for aggregating evidence and forming trust assessments.[8] Agents, representing a coalition's information providers and consumers, interact in a series of rounds. Within each round, each consumer requests information from a provider with a probability of 0.8. Each agent also will also leave the society and be replaced with a probability of 0.01 in each round, introducing a degree of dynamism. Initially, we created 500 information providers and 40 information consumers, with the latter making decisions about which providers to trust and sharing their experiences with other consumers.

Agents are generated from a number of hidden profiles that determine the agents' behavioral characteristics. Each profile is associated with two Gaussian distributions, representing the stochastic behavior given the consumer's actions $a+$ and $a-$. Here, $a+$ represents the provider choosing to meet consumer expectations and provide high-quality information, with the risk that sensitive information will be disclosed. In contrast, $a-$ represents the choice to obfuscate, redact, or otherwise provide information that is less likely to meet the consumer's expectations but entails lower risk of disclosing sensitive information.

When an information provider accepts a delegated task, it selects its preferred action, and the outcome—that is, the degree to which this information satisfies the consumer's explicit quality expectations—is drawn from the corresponding Gaussian distribution. If the contract proposed by the consumer isn't acceptable to the provider, the provider might reject it outright. For simplicity, we assume that consumers consider outcomes above 0.5 to signify success and anything lower to indicate failure.

We evaluated our mechanism in four different conditions:

- *Simple*. Consumers don't attempt to induce a desired provider choice. Instead, they pay only the minimum asking price and use only nonconditional trust models—that is, trust models that are based on observations of the final outcome only, and not on the trustee's action.
- *Unmonitored (incentive only)*. Consumers create contracts aiming to induce the provider to make the preferred choice based on conditional trust models, but monitoring isn't allowed.
- *Monitored*. The same as the unmonitored mode, but consumers can opt to create a monitored contract for a cost.
- *RI*. In this condition, the unmonitored, monitored, and RI contracts are all available.

When evaluating a set of candidates, the consumer generates a contract from each of the available types—that is, simple, monitored, and so on—for each candidate. The candidate–contract pair that yields the highest expected payoff for the consumer is proposed to that candidate.

We found that decision making and delegation using trust and controls (that is, monitoring, incentives, and RIs) can be beneficial, even when those controls are costly to implement. We evaluated our system with respect to two hypotheses: that consumers using a mixture of trust and control will outperform those who use only trust; and that agents will prefer to monitor less as time progresses, indicating an increase of trust.

We hypothesized that trustors would perform better when using control strategies and trust, rather than trust alone. Figure 3 shows our system's performance in each of the four cases above, measuring the average change in utility experienced by the society's trustors in each interaction. Error bars represent one standard deviation in the data. The simple strategy had an average utility gain of 1.8 over the course of the experiment. In the unmonitored case, the model performs very poorly, obtaining an average utility just above or below 0. This is to be expected; this strategy uses conditional trust models,

yet has no way of obtaining conditional evidence, and thus can't build conditional trust. In this way, this model acts as a control condition, showing the behavior of an agent who performs no better than chance. In the monitored mode, the model performs much better, rising to around 2.4 by 500 interactions. The RI model outperforms the others, achieving around 3.6 by 500 interactions.

We also expected that agents would become less motivated to monitor behavior as trust increases. Figure 4 plots the number of monitoring actions performed by all delegating agents as a function of time, using the full RI decision model with stereotyping. As the figure shows, there's a general trend toward fewer monitoring actions over time. Given the system's dynamism, agents occasionally lose their preferred partners, which explains why monitoring actions don't disappear entirely. However, the use of stereotypes means that even once all known agents have been replaced, stereotypical trust is sufficient to make unmonitored delegation possible. When stereotypes are disabled, monitoring actions are chosen with a higher frequency and, although it fluctuates over time, the rate remains high.

Finally, we evaluated the difference in average utility gain in the full (that is, the *RI*) model both with and without stereotyping. Figure 5 shows that using a stereotyping trust model in conjunction with monitoring and control results in a significant increase in performance. It's interesting to note that, because conditional trust models can also employ stereotypes, different stereotypes might apply to the same agent based on the actions it chooses.

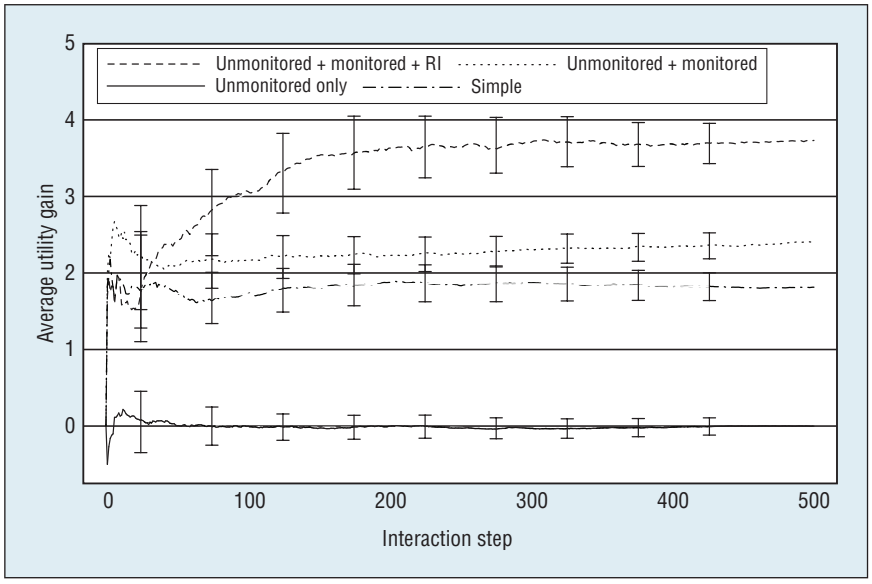

Figure 3. Performance of decision-making agents in different conditions.

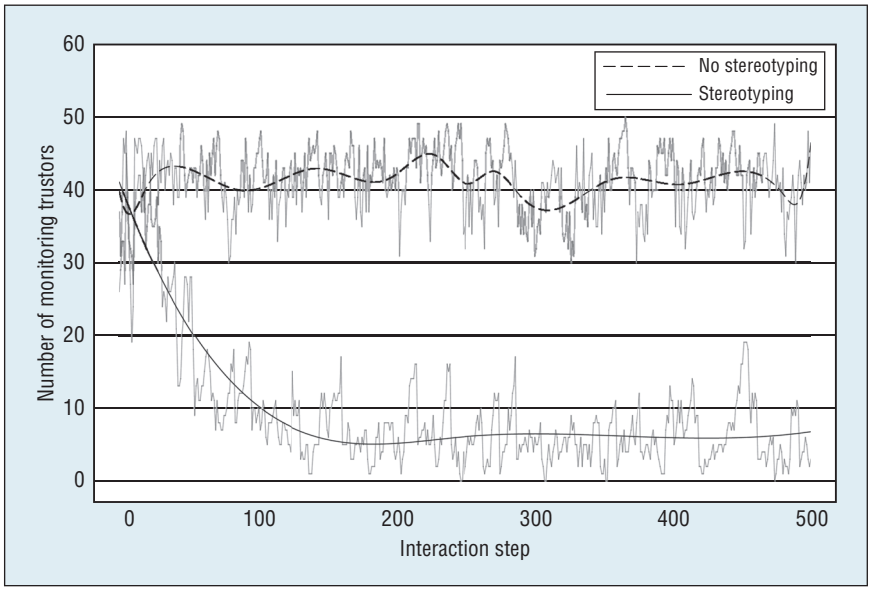

Figure 4. Monitoring instances over time.

Trust is necessary to facilitate effective collaboration in modern coalition operations. Probabilistic mechanisms provide a comprehensive trust-assessment framework that, when augmented by stereotypical assessments, can help overcome the recurring "trust bootstrapping" problem found in dynamic coalition settings. To support decision making, we propose mechanisms that supplement trust with control and consider the risks and rewards involved in a decision.

Many open problems remain in this area. Although we've focused on dyadic interactions, more complex interactions can take place within coalitions. A task might be delegated to an agent, who then *subdelegates* the task to a third, and so on. Information might be obtained from a number of sources, which is subsequently fused and summarized. This raises many new questions; for example, how should trust in a group be assessed and updated when several

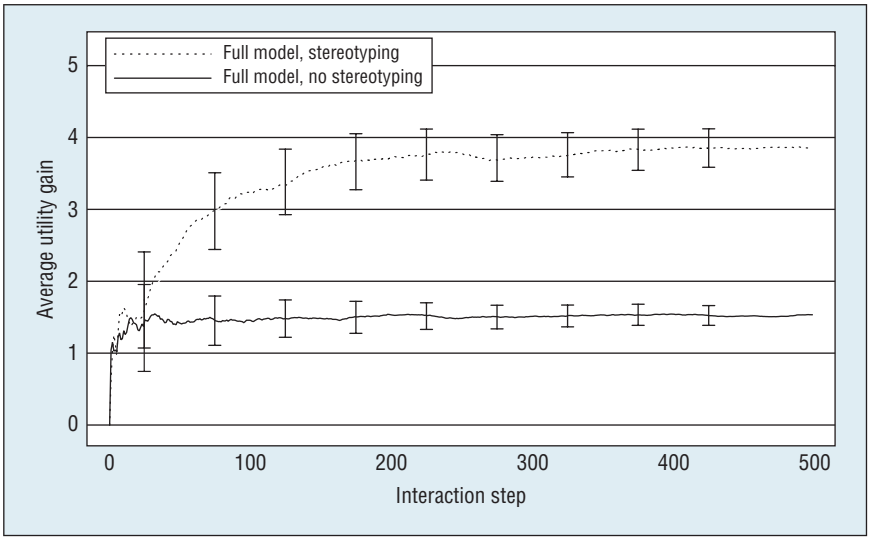

Figure 5. Performance of decision-making agents with and without stereotyping.

THE AUTHORS

Chris Burnett is a research fellow at the University of Aberdeen. His research interests include highly dynamic multiagent systems and the issues of trust that emerge in real-world applications. Burnett has a PhD in computing science from the University of Aberdeen. Contact him at cburnett@abdn.ac.uk.

Timothy J. Norman is a professor of computing science at the University of Aberdeen. His research interests include how systems might be designed and analyzed for policy compliance, and how these techniques can be applied to support human decision making; and computational models of argumentation, including dialogue strategy and formal models of delegation. Norman has a PhD in computer science from the University of London (University College). Contact him at t.j.norman@abdn.ac.uk.

Katia Sycara is a professor in the School of Computer Science at Carnegie Mellon University, where she directs the Laboratory for Agents and Semantic Technologies; she also holds the Sixth Century Chair at the University of Aberdeen (UK). Sycara has a PhD in computer science from the Georgia Institute of Technology and a Doctorate Honoris Causa from the University of the Aegean. Contact her at katia@cs.cmu.edu.

Nir Oren is a lecturer at the University of Aberdeen. His research interests include multiagent decision making, concentrating on argumentation theory, normative reasoning, and trust. Oren has a PhD in computer science from the University of Aberdeen. Contact him at n.oren@abdn.ac.uk.

agents share some degree of responsibility for a task's outcome? Implementing these mechanisms within real coalition settings certainly introduces additional challenges. For example, decision makers must be able to express their expectations, evaluate outcomes against them, and assign preferences to those outcomes.

Clearly, we must address such open problems if these systems are to inform human decision makers in dynamic, complex coalition settings. However, the mechanisms described here provide a foundation on which trust-support systems can be built, and we envisage them playing a crucial role in coalition operations in the future. ■

Acknowledgments

The US Army Research Laboratory and the UK Ministry of Defence sponsored this research under agreement W911NF-06-3-0001. The article's views and conclusions are those of the authors and shouldn't be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory, the US government, the UK Ministry of Defence, or the UK Government. The US and UK governments are authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon.

References

1. C. Burnett, T.J. Norman, and K. Sycara, "Bootstrapping Trust Evaluations through Stereotypes," *Proc. 9th Int'l Conf. Autonomous Agents and Multiagent Systems*, 2010, pp. 241–248.
2. D. Meyerson, K. Weick, and R. Kramer, "Swift Trust and Temporary Groups," *Trust in Organizations: Frontiers of Theory and Research*, Sage Publications, 1996, pp. 415–445.
3. J. Quinlan, "Learning with Continuous Classes," *Proc. 5th Australian J. Conf. Artificial Intelligence*, 1992, pp. 343–348.
4. G. Miller and A. Whitford, "Trust and Incentives in Principal-Agent Negotiations: The Insurance/Incentive Trade-Off," *J. Theoretical Politics*, vol. 14, no. 2, 2002, pp. 231–267.
5. R. Axelrod and W. Hamilton, "The Evolution of Cooperation," *Science*, vol. 211, no. 4489, 1981, pp. 1390–1396.
6. J. Borwein and A. Lewis, "*Convex Analysis and Nonlinear Optimization: Theory and Examples*," Springer Verlag, 2006.
7. C. Burnett, T.J. Norman, and K. Sycara, "Trust Decision-Making in Multi-Agent Systems," *Proc. 22nd Int'l Conf. Artificial Intelligence*, vol. 1, 2011, pp. 115–120.
8. A. Jøsang, R. Hayward, and S. Pope, "Trust Network Analysis with Subjective Logic," *Proc. 29th Australasian Computer Science Conf.*, vol. 48, 2006, pp. 85–94.