A Comprehensive Approach to Trust Management

Sandip Sen
University of Tulsa, Tulsa, OK, USA
sandip@utulsa.edu

ABSTRACT

Trust has been recognized as a key component of agent decision making in the context of multiagent systems (MASs). Though diverse trust models and mechanisms, influenced by various fields of study, have been proposed, implemented, and evaluated, we believe that the literature has ignored key aspects of pragmatic and holistic trust based reasoning. In particular, the focus of trust research has been on *a posteriori* evaluation of the trustworthiness of another agent and relatively few efforts have investigated the issue of establishment, engagement, and usage of trusted relationships. We envision that a holistic agent architecture will not use a trust module as a black-box for evaluating others but as a core component that will inform and shape interactions with other agents in the environment to best serve the decision-makers interests. Accordingly, we present a general and comprehensive trust management scheme (CTMS) that addresses key issues surrounding trust development, maintenance, and use. We present an operational definition of trust motivated by uncertainty management and utility optimization. We identify the various components required of a CTMS and their relationships and overview their use in the existing literature on trust in MAS. We welcome the MAS community to develop on the ideas presented here and build effective agent designs and implementations with fully-integrated CTMS cores.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent Systems

Keywords

trust management, engagement, trust establishment

1. INTRODUCTION

Trust plays a central and critical role in human reasoning, sustaining and promoting richness, robustness and vitality of diverse societal interactions. Trust is truly a multi-dimensional and multi-faceted, even somewhat nebulous, concept representing a somewhat loosely related set of influences on human decision-making.

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013),* Ito, Jonker, Gini, and Shehory (eds.), May 6–10, 2013, Saint Paul, Minnesota, USA.

Researchers have attempted to formalize the role of trust in multiagent interactions, and have proposed trust models that allow agents to represent, update, and use their trust in other agents and services in their environment [4, 8, 14, 15][1]. We believe that a reflective evaluation on the aspects of trust that have been relatively unexplored can identify the critical needs that can drive future trust research. Our goal is to analyze and recommend the necessary components of a CTMS that can be used by researchers to both evaluate existing trust models and develop the next-generation trust management schemes that will be more robust and effective in handling a rich, diverse, and challenging set of decision-making contexts.

To this end we consider some alternate definitions of the concept of *trust* from a computational perspective, though no one definition appears to encompass all aspects of trust in human societies as we normally perceive it. We then present our own definition of computational trust which captures the fundamental need of an agent to effectively handle uncertainty and optimize performance.

Next, we consider representative examples of real-life human and agent interactions to differentiate broad categories of trust-related decisions that autonomous entities routinely undertake. We discuss, in particular, how these decisions are correlated and introduce a holistic architecture for a CTMS module that identifies the relationships between these trust components. We also provide a generic agent architecture and discuss the integral role of the CTMS module in deciding how an agent should strategically interact with other agents in its environment.

We then review some of the well-known computational models developed by MAS researchers to identify how they match up with our proposed CTMS specifications. From this analysis we identify certain CTMS features that have been under-represented in the MAS literature.

2. TRUST AS A CONCEPT

To effectively maximize utility, a rational agent will need to coordinate, collaborate, and work with other agents. This often means that agents have to rely on other agents' decisions, e.g., that they fulfill negotiated commitments. Without any centralized authority or enforcement mechanisms in most of these open environments, commitments are non-binding. In addition, the likelihood of external offers and

[1]Due to limited space for submissions to this special track, we will only be able to cite representative literature, with omissions of important and diverse work being a regrettable and unfortunate consequence.

opportunities may provide short-term incentives to deviate from commitments. Hence, agents in open environments need to rely on distributed reputation and trust mechanisms that encourage agents to fulfill their commitments. Distributed trust schemes produce and maintain agent reputations reflecting their performance and trustworthiness and can support and sustain mutually beneficial medium to long-term relationships between self-interested agents.

Various trust definitions in literature focus either on the philosophical or pragmatic aspect of the concept [4, 8]. We use the following operational characterization that captures what it means for an agent to trust another agent:

> Trust in another agent reduces the uncertainty over that agent's independent actions which positively correlates with the truster's utility.

According to this interpretation, trust in another agent can both reduce uncertainty about outcomes and improve performance. Though this is just one of many possible useful approaches to representing and reasoning about trust and does not conceptually encompass all aspects of trust, it does provide a pragmatic framework that can guide effective and efficient trust best rational decision making.

A formal approach, though not the only one, that can be used to represent and rationally reason about and with trust information is to use a decision theoretic framework, where trust information is encoded by probabilities of likely world states and joint action outcomes that explicitly represents and reasons about other agents in a multiagent environment. From a decision makers perspective, given a set of outcomes influenced by another agent, when that agent is trustworthy, its behavior results in higher utility outcomes becoming more probable and hence results in higher expected utility. For risk neutral agents, then we can consider agents to choose actions according to the Maximum Expected Utility (MEU) principle:

$$\arg\max_{a \in A} \sum_{o \in O} Pr(o|a, M)U(o),$$

where A is the set of actions available to the agent, O is the set of outcomes possible, M is the world model of the agent and U is its utility function over outcomes.

Outcomes are determined by the joint action of the current agent and other agents in the environment and shapes the trust between them. Assuming prior knowledge of the set of actions A, the set of possible outcomes O, and the utility function, a trust model of others in the environment will then estimate the outcome probabilities, $Pr(o|a, M)$. Either a frequency based approach can be used to estimate these probabilities or one can use Bayesian priors and associated update rules for model updating. Often these outcomes will depend on unobservable parameters and may involve time dynamics. In such cases, Dynamic Bayesian networks with efficient approximate inference schemes like particle filtering may be used to estimate these probabilities.

Typically an agent develops trust estimates of another agent both from direct interactions with that agent and from trust values reported by other agents (also called *reputation*). In particular, for various reasons often cited in favor of multiagent systems, including flexibility of use, low infrastructural overhead, robustness, etc. we are interested in reputation frameworks that are distributed and peer-level rather than centralized and monolithic.

Trust is also a resource that can be leveraged to gain influence. When agent interactions are based on trust, trustworthy agents will have a larger influence on negotiated outcomes. For example, agents who are trusted to provide higher quality service may demand larger fees for their services. Trust often has to be earned at a cost. For example, a manufacturing agent may have to spend extra time and resources to meet stringent delivery deadlines when upstream suppliers delay delivery of raw materials. If, however, improved trustworthiness is rewarded with additional profitable contracts, the cost expended can be recouped many times over. In such scenarios, establishing a high reputation may be a priority for rational agents. Strategic reasoning involving trust considerations will trade-off the cost of establishing and maintaining trust in the community with the future expected profits from leveraging the trust earned.

We believe that trust is a complex, multifaceted concept and involves more than merely evaluating other's trustworthiness. A more integrated approach is necessary and should additionally address engagement of others, creating situations to evaluate trust, investing resources and time to establish your own trustworthiness, strategic use of trust information, etc. Though prior research have proposed and evaluated various trust and reputation approaches that evaluates the trustworthiness of other agents, little attention is paid to a comprehensive trust management scheme. Our proposed CTMS scheme will address trust modeling, exploration, learning, as well as both tactical and strategic reasoning to achieve the desired properties of reducing uncertainty and increasing utility.

3. COMPREHENSIVE TRUST MANAGEMENT

We believe a comprehensive trust management scheme will not only address trust evaluation, but also trust establishment and use. We now introduce an agent architecture with an embedded trust management, i.e., CTMS, module. This module stores models of other agents and runs a trust management process that interfaces both with the communication module and the decision selection mechanism. The functionalities of these sub-components are described below:

Evaluate: *The evaluation module is in charge of evaluating the trustworthiness of another agent given its history of interaction.* This is the most frequently cited and studied aspect of trust management in the literature. A *post facto* analysis of the trustworthiness of another agent is a valuable component of agent decision making.

Establish: Trust establishment is in some respect the flip side of evaluation. *This module determines the actions and the resources to be invested to establish our agent to be trustworthy to another agent.* For example, a new supplier in the market has to determine how much time, effort, and resources to allocate to process the task/contract awarded by a lucrative customer. In some sense, this module plays as critical a role in the viability of a social agent as the trust evaluation module. Unfortunately, there is very little current research that addresses this central trust issue.

Engage: *The Engage module enables rational agents to choose carefully and with strategic intention to interact and engage other agents for the purpose of evaluating their trustworthiness.* In practice, agents cannot depend primarily on accidental and circumstantial interactions to judge another agent's trustworthiness, i.e., they cannot be passive evaluators. Rather, an agent must make conscious decisions about

which other agent to interact with. In addition, agents may have to create situations and decide on task allocations that allow for trust evaluation. For example, to evaluate a new supplier in a supply chain, a company may choose to award it some contracts. The timing and importance of the contract must be carefully chosen to allow evaluation of the competence of the new supplier without jeopardizing the production or delivery schedules. The strategic creation of trust interactions that will allow for establishment or evaluation of trust is a key component of a CTMS.

Use: *This module determines how to select future courses of action based on the trust models of other agents that have been learned.* Trust considerations can influence agent decisions both in the short and in the long term. Developing trust models is key but, of necessity, they must be coupled with an effective decision procedure to utilize this knowledge. Both tactical and strategic use of trust information is key to the competitiveness of agents in open environments. In particular, careful attention must be given to the confidence in the trust values. For example, given different interaction histories with different agents, an agent must carefully balance exploitation of existing trust knowledge and investment in exploration for gathering trust models about relative newcomers in the environment.

To show the emphasis on the "evaluate" sub-module, we briefly discuss two oft-cited trust models:

FIRE: The FIRE [10] model is primarily a utility evaluation model because of the following assumptions: all agents are honest and all agents are willing to share. These two assumptions mean there is no need to utilize a central aspect of the CTMS, namely establishing trust. The current FIRE model just gathers utility information via four methods: direct experience, witness information, role based rules and third party reference. FIRE calculates a weighted mean of each of these information types and then creates a composite score. FIRE creates trust situations (engage sub-module) via a Boltzmann distribution exploration strategy. It will chose to either explore a new provider to create a trust situation or use its trust knowledge to select the provider that delivers the highest utility.

TRAVOS: TRAVOS [19], like FIRE, is primarily a trust evaluation model. It includes lying agents but does not seek to establish trust. The model does not include any strategic reasoning about if it should lie or attempt to tell the truth. TRAVOS uses direct trust and reputation to evaluate the trustworthiness of an agent. Furthermore, the TRAVOS model does not consider the utilization of the trust calculation. However, the reputation methodology is classified as creating trust situations.

4. EXISTING TRUST MODELS

Interactions in MASs might be either cooperative, competitive or simply co-existential in nature. One of the most challenging issues underlying interaction decisions in open multiagent environment is that involving trust and reputation among agents [4, 10, 7, 12, 14, 17, 22]. Therefore, a central research focus in multiagent systems involves how to learn to trust other agents and how to build and maintain reputations in an agent societies.

Castelfranchi and Falcone have argued the necessity of trust in social interactions between agents with complex mental attitudes and identified the benefits of being trusted [4, 5]. They argue that though trust necessarily entails risks for

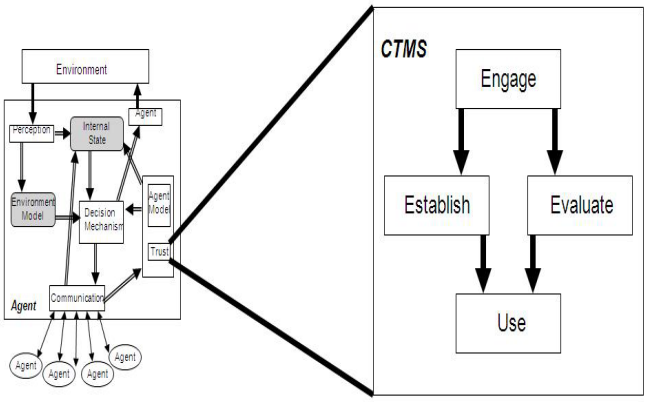

Figure 1: Principal components of the embedded trust module in an agent architecture.

delegation and collaboration, considerations of morality and use of reputation can be used to mitigate that risk. Castelfranchi, Conte, and Paolucci use normative reputation [3] to enhance the performance of agents that comply with social norms. An example of using morality to promote social relationships can be found in the SPIRE framework developed by Grosz and collaborators [9, 18]. The REGRET framework, developed by Sabater-Mir and Sierra, utilizes three dimensions of reputation: *individual dimension* based on direct interaction, *social dimension* based on feedback from social relations, and the *ontological dimension* that recognizes the multi-dimensional nature of trust and reputation [12]. Others have used generalizing from experience from similar individuals via stereotyping to aid in trust evaluations of new agents [2].

Singh and his students have also studied the management of reputation in distributed referral systems [21]. Sen and his students have studied the use of referrals to locate service providers when an agent first enters a new community with no prior knowledge of the quality of service providers or the reliability of the referrers [17]. More recent work on trust models incorporate divergent approaches including information-theoretic and fuzzy approaches to trust metrics [6, 10, 11]. In particular, there is increased focus on formal probabilistic treatment of trust with concomitant the ability to track changes in agent behaviors and performance guarantees or error bounds under realistic assumptions [20].

A significant body of work by mathematical biologists or economists on the evolution of altruistic behavior deals with the idealized problem called the Prisoner's dilemma or some other repetitive, symmetrical, and identical 'games.' To consider a well-known study in this area, Axelrod demonstrates that a simple, deterministic reciprocal scheme or the *tit-for-tat* strategy is quite robust and efficient in maximizing local utility [1]. Sen has argued that the simple reciprocative strategies are inappropriate for most real-life situations because the underlying assumptions are violated [15, 16]. Saha & Sen have studied the emergence of dominant or evolutionarily stable behaviors in such environments [13].

The research referenced above, including the ones studying learning of trust models, focus primarily on the Evaluate and Use aspects of the CTMS framework. We believe that the somewhat neglected aspects of comprehensive trust

models need sustained research focus. In particular, engagement and establishment need to be carefully studied with integrated trust based decision schemes that are informed by and, in turn, inform the usage and evaluation of trust of other agents in the environment.

5. CONCLUSIONS

We have argued for the development and use of a comprehensive trust management system as an integral component of intelligent agent architectures. We introduced a holistic conceptual view of trust decisions as reducing uncertainty and improving utility and shown that such a characterization nicely dovetails into a decision-theoretic design of a rational agent. We analyzed the requirements of an effective CTMS design and identified the corresponding fundamental modules. We reviewed some well-known trust mechanisms to highlight the current emphasis on only some of the identified CTMS modules. In particular, engagement and establishment decisions decisions are often neglected or unspecified, but can be determining factors behind the success or failure of implemented trust-based systems.

This paper begins a dialogue for a more comprehensive treatment of trust mechanisms in multiagent systems. Further investigation is needed to identify important submodules of the major trust modules identified here and probably additional modules of significance. Novel, innovative Engagement and Establishment decision mechanisms need to be developed and tested in conjunction with the Evaluate and Use modules used by existing frameworks. A holistic treatment of trust with clear identification of and contribution to the dual goals of uncertainty reduction and utility maximization should lead to principled trust module designs that can mutually inform other critical components of intelligent agent architectures. We believe that such integrated trust-based reasoning is essential for developing robust and flexible agent designs for future challenging applications.

6. REFERENCES

[1] R. Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[2] C. Burnett, N. T, and K. Sycara. Bootstrapping trust evaluations through stereotypes. In *Proceedings of the 2010 Conference on Autonomous Agents and Multi-Agent Systems, Toronto, CA*, May 2010.

[3] C. Castelfranchi, R. Conte, and M. Paolucci. Normative reputation and the costs of compliance. *Journal of Artificial Societies and Social Simulation*, 1(3), 1998.

[4] C. Castelfranchi and R. Falcone. Principles of trust for MAS: Cognitive autonomy, social importance, and quantification. In *Proceedings of the Third International Conference on Multiagent Systems*, pages 72–79, Los Alamitos, CA, 1998. IEEE Computer Society.

[5] C. Castelfranchi, R. Falcone, and F. Marzo. Being trusted in a social network: Trust as a relational capital. In *Proceedings of iTrust*, pages 19–32, 2006.

[6] R. K. Dash, S. D. Ramchurn, and N. R. Jennings. Trust-based mechanism design. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 748–755, New York, NY, 2004. ACM Pres.

[7] S. D.Ramchurn, D. Huynh, and N. R. Jennings. Trust in multiagent system. *The Knowledge Engineering Review*, 19(1):1–25, 2004.

[8] D. Gambetta. *Trust*. Basil Blackwell, Oxford, 1990.

[9] A. Glass and B. Grosz. Socially conscious decision-making. In *Proceedings of the Fourth International Conference on Autonomous Agents*, pages 217–224, New York, NY, 2000. ACM Press.

[10] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[11] S. D. Ramchurn, N. R. Jennings, C. Sierra, and L. Godo. Devising a trust model for multi-agent interactions using confidence and reputation. *Applied Artificial Intelligence*, 18(9-10):833–852, 2004.

[12] J. Sabater and C. Sierra. Reputation and social network analysis in multi-agent systems. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 475–482, New York, NY, 2002. ACM Pres.

[13] S. Saha and S. Sen. Predicting agent strategy mix of evolving populations. In *Proceedings of the Fourth Intenational Joint Conference on Autonomous Agents and Multiagent Systems*, pages 1075–1082, 2005.

[14] M. Schillo, P. Funk, and M. Rovatsos. Using trust for detecting deceiptful agents in artificial societies. *Applied Artificial Intelligence*, 14:825–848, 2000.

[15] S. Sen. Reciprocity: a foundational principle for promoting cooperative behavior among self-interested agents. In *Proceedings of the Second International Conference on Multiagent Systems*, pages 315–321, Menlo Park, CA, 1996. AAAI Press.

[16] S. Sen. Believing others: Pros and cons. *Artificial Intelligence*, 142(2):179–203, 2002.

[17] S. Sen and N. Sajja. Robustness of reputation-based trust: Boolean case. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 288–293, New York, NY, 2002. ACM Pres.

[18] D. G. Sullivan, B. Grosz, and S. Kraus. Intention reconciliation by collaborative agents. In *Proceedings of the Fourth International Conference on Multiagent Systems*, pages 293–300, Los Alamitos, CA, 2000. IEEE Computer Society.

[19] W. T. L. Teacy, J. Patel, N. R. Jennings, and M. Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Journal of Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

[20] G. Vogiatzis, I. MacGillivray, and M. Chli. A proabilistic model for trust and reputation. In *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems*, pages 225–232, 2010.

[21] P. Yolum and M. P. Singh. Engineering self-organizing referral networks for trustworthy service selection. *IEEE Transactions on System, Man, and Cybernetics*, 35(3):396–407, 2005.

[22] B. Yu and M. P. Singh. Distributed reputation management for electronic commerce. *Compuational Intelligence*, 18(4):535–549, 2002.