Combining Cognitive and Computational Concepts for Experience-Based Trust Reasoning

Eugen Staab and **Thomas Engel**
University of Luxembourg
6, rue Richard Coudenhove-Kalergi
L-1359 Luxembourg
{eugen.staab,thomas.engel}@uni.lu

Abstract

We propose a concept that combines the *cognitive* and the *computational* approaches to experience-based trust reasoning in multi-agent systems. We reemphasize that a cognitive component is vital for computationally modeling trust. At the same time, we recognize the predictive nature of trust. This suggests a combination of the two approaches. The idea is to introduce a cognitive component that produces factual good and factual bad experiences. These experiences can then be used in a predictive component to estimate the future behavior of an agent. The two components are combined in a modular way which allows the replacement of them independently. It further facilitates the integration of already existing trust update algorithms. In this work, we analyze the chain of trust processing for the concept gradually. This results in a concise survey on challenges for experience-based trust models following the proposed approach but also in general.

Introduction

Much research is directed to trust that is based on direct experiences and their propagation: *Good* experiences lead to an increase whereas *bad* experiences lead to a decrease in trust. This so called *experience-based* trust reasoning is required in open distributed systems where no centralized third party can verify an agent's trustworthiness, and therefore agents need to reason about trustworthiness of potential interaction partners on their own.

The concept for experience-based trust mechanisms that is presented in our work, builds a bridge between two different approaches to this problem. On the one hand there is the cognitive approach which was mainly promoted by Castelfranchi and Falcone (1998; 2000; 2001). In their approach the trustor tries to develop a "theory of the mind" of the trustee in order to reason about how trustworthy he will behave in future interactions. On the other hand there is the computational approach. Here, the trustor uses experiences to derive numerical values that shall represent the trustworthiness of the trustee. This process involves in the majority of cases probabilistic mechanisms (Wang & Singh 2007; Capra & Musolesi 2006; Esfandiari & Chandrasekharan 2001; Teacy *et al.* 2006) or formulas that are mainly based on intuition (Witkowski & Pitt 2000; Zacharia 1999; Sabater & Sierra 2001; Huynh, Jennings, & Shadbolt 2006).

The motivation for a fusion of the cognitive and the computational approach to trust is to seek for a model that incorporates the profundity of the cognitive approach but uses at the same time mathematically well founded mechanisms for extrapolating the experiences into the future. In our concept, the interpretation of observations gets a much higher importance than it is usual in trust reasoning. This allows us to draw the cognitive part into the component for interpretation. Consequently, the components for cognitive reasoning on one side and computational mechanisms on the other side can be strictly separated, which in turn allows a modular structure. This modular design enables an agent to tune and replace the components of our concept independently.

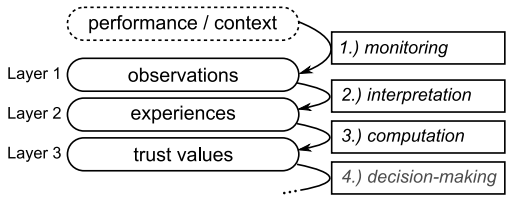

Figure 1: Modular concept for trust reasoning.

The procedural method of the paper is to analyze each component of the proposed concept by surveying the challenges it faces. Following this, we discuss how witness-information can be integrated into the concept.

Modular concept

Our concept subdivides the process of trust reasoning into four components (see also Figure 1):

1. **Monitoring:** In a first step, the performance of the trustee and his environment is monitored. This process results in a set of observations.

2. **Interpretation:** The observations obtained in step 1 are analyzed and interpreted as good or bad experiences. As we will argue in the respective section, this component requires cognitive reasoning.
3. **Computation:** The set of experiences resulting from step 2 is used as input for a computational (e.g. probabilistic) trust update algorithm.
4. **Decision-Making:** Finally (but this is outside the scope of this paper), the trust in potential interaction partners can be used as *one* criterion in decision-making. Here, game-theoretic mechanisms seem reasonable to estimate the effects a decision will have on the evolution of the system.

In the following, we explain and investigate the first three components in detail.

Step 1: Monitoring

The first step in making experiences is to *monitor* an agent's performance (or the outcome of it) and the environment within which the agent performs. Monitoring though is domain-dependent and therefore shall not be examined here in more detail. We only want to mention the most common challenge for this process which can be called *distortion* or *interference*. To quote an example, Capra & Musolesi (2006) accounted in their trust model for distortion during monitoring in pervasive systems. They assume white Gaussian noise in an agent's measurements which they statistically filter out by the use of a Kalman filter.

Even though the process of monitoring is not specific for the trust scenario, trust models must be aware of errors that can occur during this step.

Step 2: Interpretation

The monitoring of an agent's performance results in information that is neutral and thus requires an *interpretation*. Through the process of interpretation, an agent decides whether she thinks about the outcome of an interaction as good or bad experience. Note that this is the first part of subjectivity within the trust reasoning process as it is up to the interpreting agent to define what good or bad is. Now, trust is used to decide about the selection of future interaction partners. Thus, only a performance that has a positive significance for future interactions should be interpreted as good experience; and accordingly, only a performance that has a negative significance for future interactions should be interpreted as bad experience.

There are many ways to carry out an interpretation. The most common approach found in trust modeling is to compare an agent's performance to a *Service-Level-Agreements (SLA)*. An SLA lies within an n-dimensional space where each dimension represents a *service performance metrics*. For each such metrics a *service level objective* is defined (what an agent should achieve) – in the simplest case a threshold vector with n dimensions. The actual performance of an agent is evaluated in respect to these objectives. The experience can for instance be chosen to be "good", if the actual performance lies above the objectives in all dimensions and "bad" otherwise.

An evaluation using an SLA has the advantage that it is easily computable. However, it has several limitations in the context of trust. First, it does not handle the uncertainty that is inherent in observations. Secondly, it does not account for side-effects or proactive behavior. Finally, an SLA does not ask for the reasons for a divergence between the objectives and the actual performance. In particular, it does not account for the mental state of the performing agent as Castelfranchi & Falcone (2000) claim. The distinction between incompetence of an agent and malicious behavior is not possible. We therefore stress the importance of a sophisticated component for making experiences that amongst others comprises cognitive modeling. It could be argued that errors in this step of interpretation can be summed out statistically. This is not correct because already few errors together with the self amplifying property of trust (Falcone & Castelfranchi 2004) can completely change the evolution of a trust-based system.

In conclusion, the process of interpretation should, on the one hand, accurately reflect the trustor's satisfaction with the outcome of an interaction. On the other hand, interpretation should evaluate the information with respect to how the assessed agent will behave in future. A component for interpretation that incorporates both facets faces some challenges. In the following we examine these challenges one by one and refer to literature where appropriate.

Intention recognition One challenge for the step of interpretation is to distinguish between *malicious*, *selfish* and *incompetent* behavior. This distinction is essential for reasoning about the future behavior of an agent. For being able to make such a distinction, the intention of an agent has to be determined. Mao & Gratch (2004) and Demolombe & Fernandez (2005) presented two approaches to infer another agents goals and plans based on observations (the latter assume complete knowledge of the actions performed). The authors call this process *intention recognition*.

Excusableness of failures One reason for interpreting a failure of an agent falsely as bad experience is *excusableness*. Staab & Engel (2007) propose to call a failure *excusable* in respect to a *reference set* of agents, and within a time-frame $[t_1, t_2]$, iff all of the following three conditions are fulfilled:

1. The environment interfered with the performance of the task.
2. All other agents from the *reference set* would have failed to perform the task under the same conditions.
3. The performing agent tried its "best" within $[t_1, t_2]$ to accomplish the task.

The social facet of this problem is covered by the "social credit assigning problem", analyzed by Mao & Gratch (2003). They presented a computational approach to find out which agent is socially responsible to what extent for an action that she executed. They follow Weiner (1995) and use (1) the hierarchical structure of *coercion* to identify the responsible agents and (2) *intention* and *foreseeability* to determine the "intensity" of the responsibility.

Delegation of a delegation Assume, an agent α delegates a task τ to an agent β; however, β is busy and decides to delegate τ further to another agent γ. Now assume that γ fails (and the failure is not excusable). The question is, whether α should trust less in β and/or in γ for performing a similar task τ' in the future; or whether α should trust β less in general. In other words, some clarification is needed of how the trustworthiness of both agents β and γ towards α is involved in the process of "multiple delegation". To the knowledge of the authors, this problem has not yet been discussed in the area of trust research.

Side-effects *Side-effects* that the performance of an agent causes, can influence its trustworthiness. Assume for example, an agent β successfully accomplishes a task τ in favor of agent α. However, agent β's actions produce side-effects that have a negative impact on agent α. Then, α should trust β more in how close she sticks to a deal (good experience). At the same time, α should trust less in β's way to perform (bad experience). So, the interaction produces several experiences.

Note that side-effects can appear heavily delayed which can raise two problems. First, over time it gets harder and harder to establish a relationship between the performance of agent β and the side-effects. Second, they are possibly not yet observed when trust reasoning is done (e.g. in the case of *global warming*).

Proactive behavior Analogously to side-effects, *proactive behavior* should be considered in interpretation. Assume an agent α asks agent β to accomplish task τ_1; but β decides to *proactively* accomplish task τ_2 as she believes, that in this case the resulting state is more desirable for α than in the case of τ_1. Agent β succeeds and his belief turns out to be true. Agent α can be unhappy about the fact that β did not exactly do what he asked for; in this sense it is a bad experience and the according trust is diminished. But he is happy about β's intelligence and it is a good experience in this sense. As in the case of side-effects, we get several experiences for only one interaction here.

Incomplete information Monitored data can be incomplete or distorted and therefore experiences can comprise a certain amount of *uncertainty*. If possible, mechanisms should be provided that help to reduce this uncertainty. For example, whenever only the outcome of a performance but not the performance itself can be perceived, an agent can reason about the causes for the outcome. The huge area of research concerned with this problem is called causality (e.g. see Pearl, 2000). To show its practical relevance in the context of trust, we give a concrete example in the area of networking. Assume agent α asks β to upload a signed data file to a server. As it is problematic to observe β's actions, α can try to download the file using another route; if this is possible and the file can be verified, β must have succeeded.

If it is impossible to reduce the uncertainty of interpretation, it is propagated to the step of computation.

Step 3: Computation

Given a set of experiences, a trust update algorithm tries to extrapolate the data into the future. The algorithm outputs some measure, usually a trust-value – a number in a continuous or discrete interval –, that represents a positive or negative forecast for future interactions with the trustee. For prediction, for instance *time series methods* (e.g. *trend estimation*), *econometric methods* (e.g. *regression analysis*) or *probabilistic forecasting* can be used. Subjectivity is present also in this step: Parameters of the model can determine whether the agent is optimistic or pessimistic, how unforgiving he is (i.e. how heavy old experiences should be weighted with a so called *aging factor*), etc.

In the following, we concisely survey the challenges that are specific to these algorithms.

First time offender problem Think of an agent who pretends trustworthiness in many interactions, and as soon as other agents have developed a strong trust in her, she behaves maliciously. This problem is called the *first time offender problem* (Rehák & Pechoucek 2007). In one extreme case, experiences accurately reflect intentions. Then, the first time offender problem is irrelevant. The other extreme case is that experiences ignore intentions. Then this problem cannot be tackled by any trust update algorithm.

Changes in character The first time offender problem is about a change from a feigned character to the real character. Opposed to that, the real character of an agent can change, too. If a character changes gradually, trust models can for instance use linear prediction techniques. If a character changes suddenly, this can for a certain time not be distinguished from statistical outliers, and thus it takes some time to recognize it.

Change of identity In open systems, agents can freely leave the system and reenter with a new identity. This enables agents, once they are not much trusted anymore by other agents, to simply change their identity. *Identity recognition* helps to prevent this to happen (Rehák & Pechoucek 2007): The identity of an

agent is recognized on the basis of certain characteristics which are assumed to be the same after leaving and reentering a system. An example for such a characteristic would be the sending characteristics of a wireless network card (Stefano *et al.* 2006). Identity recognition helps moreover to avoid that the first time offender problem can be exploited more than once by an agent.

Undecidability Trust reasoning that is based on experiences has a fundamental limit. This limit becomes evident in the context of *confidentiality* which is illustrated by the following example. Assume, agent α transfers confidential information to agent β and both reach a non-disclosure agreement. In the case that α detects a violation of the agreement by β, the respective indications or observations can be interpreted as bad experience. In the case that α detects nothing, α has no experiences and cannot decide about β's trustworthiness. The same problem exists also whenever tasks are delegated without time limit. At no point in time it is known whether the agent will or will not accomplish the task (only in case of accomplishment). Basically, we face the *halting problem* (Sipser 1996) or the problem of verifying a scientific theory.

Context-sensitivity Researchers agree on the property of trust which is called *context-sensitivity*. It describes the fact that experiences in a certain context are only significant for trust-based decisions in similar contexts. Therefore, trust values should always be associated to a specific context. In literature, the term "context" is used in two different ways. Sometimes "context" refers to the type of task that is to be performed (Grandison & Sloman 2003; Kinateder, Baschny, & Rothermel 2005). McKnight & Chervany (1996) give the example that "one would trust one's doctor to diagnose and treat one's illness, but would generally not trust the doctor to fly one on a commercial airplane". And sometimes the term "context" refers to the conditions under which a task is performed (Şensoy & Yolum 2007; Rehák & Pechoucek 2007). The first notion refers only to the internal causes of an agent's performance (Falcone & Castelfranchi 2004) whereas the second notion refers to an interplay of internal and external causes. It is obvious that the two meanings are relevant for trust and so a trust-model should account for both of them – but in different ways.

Incomplete information Although some experiences with full certainty can be given, an extrapolation of these experiences will always remain uncertain. This is especially the case when the conflict within the experiences is high (Wang & Singh 2007). Therefore a distinction of the certainty of *experiences* and *trust values* is necessary.

Integration of Witness-Information

In this section we discuss how the exchange of witness-information fits in our concept. Witness-information is any information which is exchanged between two trustors and which concerns the trustworthiness of a trustee. This information helps trustors to overcome the stage where no direct experiences with a certain trustee could be made yet, and lets agents base their "trust" reasoning on a larger knowledge base. We want to stress that reasoning, which is only based on witness-information, is about *reputation* and not about trust.

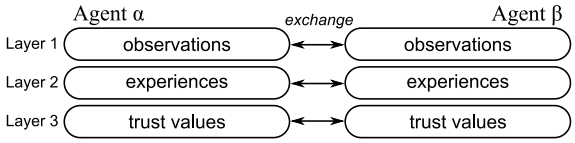

Figure 2: Exchange of witness-information.

As Figure 2 reveals, our concept allows the exchange of information on all three layers of the trust reasoning process. It holds that the higher the layer for information exchange is, the lower is the level of subjectivity. This is the reason why Şensoy & Yolum (2007) exchange witness-information on layer 1. If information is exchanged on layer 2, a compliant interpretation is necessary but the subjectivity for the forecast using this information is still preserved. Mechanisms that exchange information on layer 3 also loose this kind of subjectivity. It is dangerous to exchange information on layer 2 or 3 without knowing about the witness's subjectivity. Therefore, Abdul-Rahman & Hailes (2000) introduced the concept of "semantic distance" between two agents that exchange information on layer 3.

For the sake of completeness we want to mention that in order to use witness-information on any layer, mechanisms are required that detect inaccurate information sources. Inaccuracy can result from incompetent but also lying agents. Approaches to this problem can be found in literature (Teacy *et al.* 2006; Huynh, Jennings, & Shadbolt 2006; Whitby, Jøsang, & Indulska 2004).

Future Work

Motivated by this work, we plan to develop a formal trust model that uses a more refined mechanism for incorporating experiences. An experience will be represented as a triple $\langle s, i, u \rangle$ where s measures the satisfaction of the trustor, i reflects the intention of the trustee (good or bad) and u represents the uncertainty for the experience. Several experiences will be possible for only one interaction. Further, we will allow the exchange of witness-information on all three layers: As soon as two agents sense similar transitions from layer i to layer $i+1$, they can start exchanging information on layer $i+1$ (which means less computation).

Conclusion

We presented a modular concept for experience-based trust reasoning that combines the cognitive and the computational approach. By analyzing each component of the concept, we surveyed challenges and limits that are specific to trust reasoning that bases on experiences. Finally, we showed how the use of witness-information fits in our approach. The proposed concept and the overview on challenges is intended to be used as basis for developing sophisticated trust models that incorporate direct experiences and witness-information.

References

Abdul-Rahman, A., and Hailes, S. 2000. Supporting trust in virtual communities. In *Proc. of the 33rd Hawaii Int. Conf. on System Sciences (HICSS '00)*. IEEE.

Capra, L., and Musolesi, M. 2006. Autonomic trust prediction for pervasive systems. In *Proc. of the 20th Int. Conf. on Advanced Information Networking and Applications - Volume 2 (AINA '06)*, 481–488. IEEE.

Castelfranchi, C., and Falcone, R. 1998. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proc. of the 3rd Int. Conference on Multi-Agent Systems (ICMAS '98)*, 72–79. IEEE.

Castelfranchi, C., and Falcone, R. 2000. Trust is much more than subjective probability: Mental components and sources of trust. In *Proc. of the 33rd Hawaii Int. Conf. on System Sciences (HICSS '00)*. IEEE.

Şensoy, M., and Yolum, P. 2007. Ontology-based service representation and selection. *IEEE Trans. Knowl. Data Eng.* 19(8):1102–1115.

Demolombe, R., and Fernandez, A. M. O. 2005. Intention recognition in the situation calculus and probability theory frameworks. In *Proc. of the 6th Int. Workshop on Computational Logic in Multi-Agent Systems (CLIMA '05)*, 358–372. Springer.

Esfandiari, B., and Chandrasekharan, S. 2001. On how agents make friends: Mechanisms for trust acquisition. In *Proc. of the 4th Workshop on Deception, Fraud and Trust in Agent Societies*, 27–34.

Falcone, R., and Castelfranchi, C. 2001. Social trust: a cognitive approach. 55–90.

Falcone, R., and Castelfranchi, C. 2004. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Proc. of the 3rd Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS '04)*, 740–747. IEEE.

Grandison, T., and Sloman, M. 2003. Trust management tools for internet applications. In *Proc. of the 1st Int. Conf. on Trust Management (iTrust '03)*, 91–107. Springer.

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

Kinateder, M.; Baschny, E.; and Rothermel, K. 2005. Towards a generic trust model - comparison of various trust update algorithms. In *Proc. of the 3rd Int. Conf. on Trust Management (iTrust '05)*, 177–192. Springer.

Mao, W., and Gratch, J. 2003. The social credit assignment problem. In *Proc. of the 4th Int. Conf. on Intelligent Virtual Agents (IVA '03)*, 39–47. Springer.

Mao, W., and Gratch, J. 2004. A utility-based approach to intention recognition. In *Proc. of Workshop on Agent Tracking (at AAMAS '04)*.

McKnight, D. H., and Chervany, N. L. 1996. The meanings of trust. Technical report, University of Minnesota.

Pearl, J. 2000. *Causality: models, reasoning, and inference*. Cambridge University Press.

Rehák, M., and Pechoucek, M. 2007. Trust modeling with context representation and generalized identities. In *Proc. of the 11th Int. Workshop on Cooperative Information Agents (CIA '07)*, 298–312. Springer.

Sabater, J., and Sierra, C. 2001. REGRET: Reputation in gregarious societies. In *Proc. of the 4th Workshop on Deception, Fraud and Trust in Agent Societies*, 194–195. ACM.

Sipser, M. 1996. *Introduction to the Theory of Computation*. Int. Thomson Publishing.

Staab, E., and Engel, T. 2007. Formalizing excusableness of failures in multi-agent systems. In *Proc. of the 10th Pacific Rim Int. Workshop on Multi-Agents (PRIMA '07)*. Springer. to appear.

Stefano, A. D.; Terrazzino, G.; Scalia, L.; Tinnirello, I.; Bianchi, G.; and Giaconia, C. 2006. An experimental testbed and methodology for characterizing IEEE 802.11 network cards. In *Proc. of the 2006 Int. Symp. on a World of Wireless, Mobile and Multimedia Networks (WOWMOM '06)*, 513–518. IEEE.

Teacy, W. T. L.; 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 *Proc. of the 20th Int. Joint Conf. on Artificial Intelligence (IJCAI '07)*, 1551–1556.

Weiner, B. 1995. *The Judgment of Responsibility*. Guilford Press.

Whitby, A.; Jøsang, A.; and Indulska, J. 2004. Filtering out unfair ratings in bayesian reputation systems. In *Proc. of the 7th Int. Workshop on Trust in Agent Societies (at AAMAS '04)*.

Witkowski, M., and Pitt, J. 2000. Objective trust-based agents: Trust and trustworthiness in a multi-agent trading society. In *Proc. of the 4th Int. Conf. on Multi-Agent Systems (ICMAS '00)*, 463–464. IEEE.

Zacharia, G. 1999. Collaborative reputation mechanisms for online communities. Master's thesis, Massachusetts Institute of Technology.