Multimodal Trust Formation with Uninformed Cognitive Maps (UnCM)

(Extended Abstract)

Michele Piunti
Reply Whitehall
Rome, Italy
m.piunti@reply.it

Matteo Venanzi
AIC Group, Electronics and Computer Science
University of Southampton
m.venanzi@ecs.soton.ac.uk

Rino Falcone, Cristiano Castelfranchi
Institute of Cognitive Sciences and Technologies, ISTC-CNR
{r.falcone,c.castelfranchi}@istc.cnr.it

ABSTRACT

This work describes a cognitive heuristic allowing agents to assess trust and delegations merging heterogenous information sources. The model is realized through Uninformed Cognitive Maps, based on the combination of: (*i*) categorization abilities (*ii*) history of personal experiences (*iii*) context awareness.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent agents, multiagent systems*

General Terms

Algorithms, Performance, Design, Theory

Keywords

Cognitive models, Trust, Social Systems, Social simulation

1. MULTIMODAL TRUST FORMATION

Crucial abilities for agents engaged in open systems is to decide how to coordinate activities and whether (or not) delegate tasks to other, possibly unknown, agents. Trust based interactions have been proposed as a suitable model to achieve such a subjective coordination. But, placed in the context of open and dynamic systems, the main issue of trust management is a problem *trust formation*. Existing approaches to trust formation refer to *subjective experiences* and *reputation* mainly. Subjective experiences are typically exploited in evaluating the outcomes of previous transactions, and therefore they are limited by the need of multiple and repeated interactions between the same agents. Reputational approaches have been proposed to establish trustworthy interactions with possibly unknown counterparts [7, 5]. The downside is the need of a network of reputation providers, being each reputational information possibly biased or corrupted. Other approaches push on the multifaceted relationship between environments, context awareness and trust management. Finally, the relevance of categories for trusting strangers has been remarked in the work of Falcone et al [4]. Categorial reasoning for trust formation has also been recently explored by Burnett et al. [2]. In their work, they propose the notion of stereotypical trust (stereotrust) as a categorial prejudice that agents may take into account in order to assess trust *in absence* of direct evidences. The mechanism adopts data mining techniques applied over the database of past transactions.

The approach proposed in this research aims at combining three different information sources into a unique reasoning process. Multimodal trust formation is realized through a novel mechanism called *Uninformed Cognitive Map* (UnCM), where the introduction of learning mechanisms further allows to establish a series of emergent relations between a rich set of information sources and the trustworthiness of unknown trustees. In doing so, we rely on the socio cognitive theory of trust [3], according to which trust is grounded on detectable cognitive ingredients.

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

2. UNINFORMED COGNITIVE MAPS

In cognitive agents, the problem of trust formation can be translated in the problem of retrieving the constituent beliefs of trust. Cognitive trust is treated as a relational construct between a trustor (trust giver, ag_i) and a trustee (trust receiver, ag_j) which can be established in a given environment/context E, and about a defined task to be fulfilled (τ): $Trust(ag_i, ag_j, E, \tau)$. Trust is then *graded* over multiple dimensions. The degree of trust (DoT) comes from a series of cognitive primitives, which can be summarized in terms of trustor's beliefs and goals. The approach takes into account the three contributions that play a crucial role in trust formation: *Bel*($Can_{ag_j}(\tau)$), that is trustor believes that ag_j is potentially able to fulfill τ (i.e., ag_j has the skills, the competences, the necessary instruments for realizing that task τ); *Bel*($Will_{ag_j}(\tau)$), that is trustor believes that ag_j is potentially willing and persistent in fulfilling τ (i.e., ag_j has the motivational attitudes sufficient to perform the task τ); *Bel*($ExtFact_{ag_j}(\tau)$), that is trustor believes that the external conditions are not preventing the execution of τ by ag_j (or even: ag_i believes that ag_j will perform the task τ in an environment presenting positive or negative interferences to ag_j's behavior in order to achieve the task τ). Summing up, an agent ag_i trusts ag_j about a task τ and in the conditions E, if ag_i's DoT overcomes a given threshold σ: $DoT_{ag_j,E,\tau} > \sigma$. The model resembles the notion of *Krypta* and *Manifesta*, according to which agents' manifesta are signals, or observable traces, recalling agents' krypta, which are the internal properties (*qualities*, *abilities* or *powers*) finally determining agents' behaviors on specific tasks

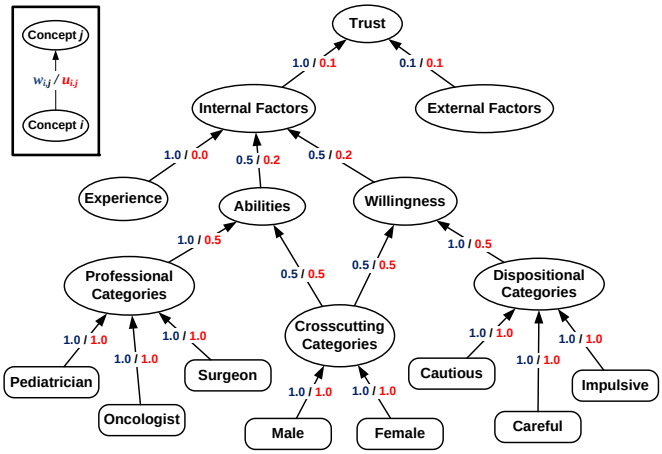

Figure 1: UnCM implementing the socio-cognitive trust model with multiple dimensions.

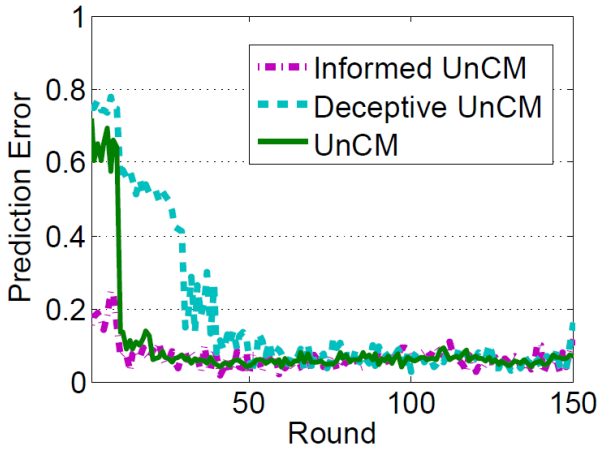

Figure 2: Plot of the prediction error for the UnCM, Informed UnCM, Deceptive UnCM over 150 simulation rounds.

or contexts [1].

Uninformed Cognitive Maps (UnCM) are a novel approach hybridizing cognitive modeling and learning. They are based as an extension of Fuzzy Cognitive Maps, a computing technique successfully applied in several domains for modeling knowledge-based systems [6]. An UnCM is as a graph modeling causal processes by means of concepts and causal relations placed on different dimensions (Fig. 1). The UnCM layout is designed by domain experts using an off-line setting. At design time, the relevant concepts of a problem domain are identified, and their reciprocal influences are quantitatively modeled by weighted connections. The causal impact between two concepts A_i and A_j is then measured by the weight of the connection $w_{i,j}$, taken in the interval $[-1, 1]$.

3. EXPERIMENTAL EVALUATION

Evaluation concerned a simulated agent society in a medical domain, with 100 trustees having krypta randomly selected from a repository of 2500 profiles. Every profile is characterized by three types of categories, which can be of *professional, dispositional* or *crosscutting*. The experiment discussed here used the *pneumonia* task, for which the best categorial profile is assumed to be $\langle$*pediatrician, cautious, female*$\rangle$. The outcome of trustee execution is referred in terms of *score*, while the accuracy of trust formation is measured in terms of *prediction error* as the distance of the predicted DoT from the real delegation outcome: $error = |DoT_{ag_j} - score|$. Setting also takes into account the environmental influences, defined as a ρ parameter indicating the contribution of situated conditions to the executor's performance. Hence, each task execution may receive an influence randomly distributed in the range $[-\rho, +\rho]$ System openness is determined by the parameter δ, which determines the number of trustees replaced at each round. Finally, L sets the interval rounds after which the trustors update their learning model over the experiences history. The model has beens compared with well-established approaches to data analysis and decision making, as neural networks (Neural agents) and agents using stereotypes and data mining mechanisms (Stereotrust agents). Experiments pointed out the abilities of UnCM strategies to perform task delegation based on multimodal trust attribution. Either context awareness and experiences play a pivotal role in trust formation in open and dynamic systems. The adopted UnCM, in particular, allows to learn to which extent the single categories fit for a given tasks, thus drastically enhancing delegation-making.Fig. 2 shows the performance of UnCM in minimizing errors: experiments show that categorial evidences *emerge* with respect to the ongoing tasks—also *without* requiring any initial categorial knowledge. The mechanism manages in a unique function heterogeneous information sources, ranging from personal experiences, to manifesta and external influences. Thanks to the UnCM learning algorithm, categories are revised, or devised from scratch, and the categorial information is combined to personal experiences and environmental conditions encountered. Differently from Neural and Stereotrust agents , the UnCM agents are also able to maintain a meaningful *semantic* of influences between concepts and their connections. Influences of the single categories on a given task represent a key aspect and, using UnCM this information is explicitly readable and updated *online*. Limitations of the current approach pave the way to future work. To evaluate the scalability of the proposed approach, applications in different domain as social networks will be devised.

4. REFERENCES

[1] M. Bacharach and D. Gambetta. Trust as Type Detection. In *Trust and deception in virtual societies*, 2001.

[2] C. Burnett, T. Norman, and K. Sycara. Bootstrapping Trust Evaluations through Stereotypes. In *Autonomous Agents and Multiagent Systems (AAMAS-10)*, pages 241–248, 2010.

[3] C. Castelfranchi and R. Falcone. *Trust Theory. A Socio-Cognitive and Computational Model*. Wiley, 2010.

[4] R. Falcone, M. Piunti, M. Venanzi, and C. Castelfranchi. From manifesta to krypta: The relevance of categories for trusting others. *ACM Transactions on Intelligent Systems and Technology.*, 2011.

[5] T. G. Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated Trust and Reputation model for Open Multi-Agent Systems. *Journal of Autonomous Agent and Multi-Agent Systems*, 13:119–154, 2006.

[6] B. Kosko and J. Burgess. Neural Networks and Fuzzy Systems. *The Journal of the Acoustical Society of America*, 103:3131, 1998.

[7] B. Yu and M. P. Singh. An evidential model of distributed reputation management. In *Autonomous Agents and Multiagent Systems (AAMAS-02)*, pages 294–301, 2002.