Facing Openness with Socio-Cognitive Trust and Categories

Matteo Venanzi[1,2], **Michele Piunti**[1], **Rino Falcone**[1] **and Cristiano Castelfranchi**[1]
[1] Institute of Cognitive Sciences and Technologies, ISTC-CNR, Rome, Italy
{michele.piunti,rino.falcone,cristiano.castelfranchi}@istc.cnr.it
[2] School of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK
mv1g10@ecs.soton.ac.uk

Abstract

Typical solutions for agents assessing trust relies on the circulation of information on the individual level, i.e. reputational images, subjective experiences, statistical analysis, etc. This work presents an alternative approach, inspired to the cognitive heuristics enabling humans to reason at a categorial level. The approach is envisaged as a crucial ability for agents in order to: (*1*) estimate trustworthiness of unknown trustees based on an ascribed membership to categories; (*2*) learn a series of emergent relations between trustees observable properties and their effective abilities to fulfill tasks in situated conditions. On such a basis, categorization is provided to recognize signs (*Manifesta*) through which hidden capabilities (*Kripta*) can be inferred. Learning is provided to refine reasoning attitudes needed to ascribe tasks to categories. A series of architectures combining categorization abilities, individual experiences and context awareness are evaluated and compared in simulated experiments.

1 Introduction

Open agent systems are characterized by heterogenous agents with their own goals which are assumed to join and leave the system. This produces multiple interactions each with possibly uncertain context conditions and unpredictable outcomes. Crucial abilities for agents are in this case to autonomously decide how to coordinate activities and to delegate or not tasks to other agents. Trust based interactions have been proposed as a suitable model in order to effectively achieve goals jointly: agents capable to assess trustworthy interactions have better chances to reduce the risk of failures and to promote desirable outcomes. The problem of trust management is, for agents, a problem of *trust formation*. Indeed, assessing trust leads agents to the need to analyze multiple information sources along several dimensions. Two main approaches to trust formation exist: they refer to *subjective experiences* and *reputation* respectively, and their effectiveness presents a series of benefits and drawbacks as showed for instance in [Fullam and Barber, 2007]. Subjective experiences are typically exploited in evaluating the outcomes of previous transactions [Littman and Stone, 2001], but they are limited by the need of multiple and repeated interactions with the same agents. Besides, reputational reports are exploited to establish trustworthy interactions with possibly unknown counterparts [Yu and Singh, 2002; Sabater, 2003; Huynh *et al.*, 2006; Fullam and Barber, 2007; Burnett *et al.*, 2010]. The downside here is the need of a network of reputation providers, being each reputational information possibly corrupted by biased reporters.

This work refunds the problem of trust formation in open and dynamic systems by combining subjective experiences and reputation with the use of abstract categories and categorial reasoning. The proposed approach is inspired by a heuristics commonly exploited by humans, that is the cognitive ability to represent the society through categories of individuals grouped by common features and capabilities. Thanks to a process of ascription, considering a unknown agent as belonging to a known category allows a cognitive agent to infer (or at least attribute) specific internal features for such unknown agent. On this basis, to determine how that agent will perform in specific situations. In this sense, the model recalls the notions of *Krypta* and *Manifesta* [Bacharach and Gambetta, 2001], 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 and contexts. The aim of this work is also to exploit multiple information sources to inform trust formation. Categorial reasoning is exploited in order to infer hidden information from observable features, thus easing the delegation process as well as the decisions needed to establish trustworthy interactions. At the same time, subjective knowledge and reputation (coming from subjective or collective experiences, as in [Littman and Stone, 2001; Yu and Singh, 2002]) as well as context awareness (treated by appraising situated conditions, as in [Tavakolifard *et al.*, 2008]) are also considered in a coherent computational model. In doing so, the paper proposes task delegation as a trust based decision making, integrating *individual*, *categorial* and *situated* reasoning. Section 2 describes categorial reasoning in a cognitive approach to trust; Section 3 formalizes the problem domain, describing agent abstract architecture; Section 4 presents different approaches for categorial reasoning; Section 5 evaluates the models with comparative experiments; finally, Section 6 concludes the paper with final remarks and perspectives.

2 Trusting Categories of Agents

The suggestion to exploit categorial knowledge to assess trust in open systems is not new, and it has been theoretically explored for ascertaining beforehand the trustworthiness of possible unknown counterparts [Barber, 1983]. In the context of computational models, the work by Wojcik et al. introduced the notion of prejudice filters to perceive particular trustees attributes [2006]. Rules are extracted to avoid distrusted interactions, thus denying transactions which may be expected as not profitable. The Stereotrust approach proposed by Brunett et al. allows agents to build stereotypes based on the analysis of outcomes of past interactions [2010]. This mechanism is inspired by data mining techniques to be applied by trustors over the database of past interactions. Categories emerge as clusters of agents, grouped by similar transaction outcomes.

The approach proposed in this paper relies on the more general theory of socio-cognitive trust [Castelfranchi and Falcone, 2010], according to which trust is a notion appraised by agents in terms of cognitive ingredients. Cognitive trust is treated as a relational construct between a trustor (ag_i) and a trustee (ag_j) which can be established in a given environment/context E, and, most important, about a defined activity or task to be fulfilled (τ): $Trust(ag_i, ag_j, E, \tau)$. Trust is then a *graded* construct which trustors (trust givers) ascribe to trustees (trust receivers). The degree of trust (DoT) comes from a series of cognitive primitives, which can be summarized in terms of beliefs and goals. Summing up, an agent ag_i trusts ag_j about a task τ and in the conditions E, if DoT overcomes a given threshold σ:

$$DoT_{ag_j,E,\tau} > \sigma$$

In a group of possible trustees, we assume the trustor will prefer the one having the higher DoT.

Besides the personal level of analysis related to the features of a single, known trustee, the socio cognitive model includes the use of categories to estimate the behavior of a group of unknown trustees. We here refer to categorial reasoning as the heuristic adopted by cognitive agents (*i*) to recognize to which extent a given category is suitable for a given task, and (*ii*) to delegate the task to the trustee better fitting the ascribed category. The categories to which a given trustee belongs can be inferred on the basis of a set of information readable over agents' observable features. Such observable properties are the *signs* indicating trustees' unobservable properties, namely the internal states which finally determines their actual behavior. In doing so, the overall informational structure can be divided in observable and non observable (*manifesta* and *krypta* according the model in [Bacharach and Gambetta, 2001]). However, in assigning a category to a given trustee, the trustor is assumed to ignore the trustee's real krypta.

3 Reasoning with Categories

In order to provide a practical taste of the approach, in what follows we refer to categories, tasks, activities inspired to the medical domain.

Categories are assumed as belonging to three different behavioral attitudes. *Professional* categories refer to the pragmatic abilities to bring about goals in a given task. Examples

Chickenpox	
Abilities	
pediatr_spec	99
manual	90
literature	80
technique	90
Dispositions	
availability	90
caution	80
attention	70
Cross	
female	*true*

Pediatrician	
Professional	
pediatr_spec:	[99 . . . 100]
manual:	[70 . . . 100]
literature:	[60 . . . 100]
technique:	[70 . . . 100]

Careful	
Dispositional	
caution:	[80 . . . 100]
attention:	[90 . . . 100]
availability:	[40 . . . 60]

Female
Crosscutting

Table 1: Example of task (left) and categories (right).

of these are *pediatrician, surgeon, dentist, oncologist*, etc. *Dispositional* categories refer to the attitudes of willingness in performing activities: Examples are *cautious, careful, impulsive,* etc. Finally, *Crosscutting* categories consider aspects which can not be considered in the above mentioned ones, for instance being *male, female, religious, atheist*, etc. Each category is specified by a record of *features*, shaped on concrete dimensions and range of values.

We model a set of tasks $\tau \in \mathcal{T}$, each task being identified by a couple (*action, goal*) where the goal is associated to a particular action to be fulfilled. Tasks are specified by a list of practical requirements that agents performing the related actions have to comply with in terms of categories.

Table 1 shows an example of representation for tasks and categories as they will be adopted in the rest of this work. The task specification includes the list of professional, dispositional, crosscutting requirements needed by agents to fulfill it. These requirements are identified by threshold values to which agents abilities must comply with. For instance, in order to achieve the *chickenpox* task, a trustee must have internal abilities (krypta) greater than the ones specified in the task (Table 1 left). Category specifications are identified by a set of features and the relative range of values. Agents belonging to a given category can be assumed to have krypta in the specified range, for instance a *pediatrician* is supposed to have a *manual_ability* between 70 and 100, a *pediatr_spec* between 99 and 100, and so on.

Each professional category, according to the modeled problem domain, is shaped on the requirements specified for a given task (see Table 1). For instance, the *Pediatrician* category is related to the *chickenpox* task by means of the *pediatr_spec* requirement. The outcome resulting from the execution of a given task is calculated as a function of the actual krypta owned by the executor agent. Krypta are compared with the thresholds specified by the task. The fulfillment value on the task is calculated using a simple matchmaking function—omitted here for simplicity.

Alg. 1 shows the abstract script of the categorial reasoning for an agent assessing trust. Given a task $\tau \in \mathcal{T}$, trustor's goal consists in delegating the task after having perceived the environmental conditions E. A configuration phase is possibly exploited through the config function, which is assumed to feed the categorizer module ϕ_τ with the trustee manifesta $mnf(ag_j)$, analysis of past interactions $\mathcal{E}xp$ and

Algorithm 1 Trustor abstract delegation process

Agent Internal State :

$\mathcal{O}thers$: Belief set storing the potential trustees actually in the MAS.
$\mathcal{E}xp$: Belief set storing experiences of past delegations.

procedure $delegate(\tau, E)$

1: **for each** $(ag_j \in \mathcal{O}thers)$ **do**
2: $\phi_\tau = \mathsf{config}_\tau(mnf(ag_j), \mathcal{E}xp, E)$
3: $DoT_{ag_j,\tau} = \mathsf{trust\text{-}eval}(\phi_\tau)$
4: **if** $(DoT_{ag_j,\tau} \geq \sigma)$ **then**
5: $L_{trustees} = L_{trustees} \cup \langle ag_j, DoT_{ag_j,\tau} \rangle$
6: **end if**
7: **end for**
8: $trustee\_agent = \arg\max_{ag_j}(DoT_{ag_j,\tau} \in L_{trustees})$
9: send$(trustee_agent, \mathtt{achieve}, \tau)$

situated conditions E. The degree of trust DoT is calculated for each agent ag_j present in the system (line 3). If DoT_{ag_j} overcomes a given threshold σ, ag_j is added to the list of possible trustees (line 5). Finally, the trustee with the highest DoT is chosen for delegation (lines 8-9). Once the delegation outcome is received, the result is added to the history of individual experiences $\mathcal{E}xp$.

The crucial element for trust formation is represented by the trust-eval function (line 3): different models could be approached in practice to implement such a function, taking into account categorical, subjective and environmental information sources.

4 Approaches for Categorial Trust

In the previous sections, we defined the categorization as key capabilities for a trustor operating in open and dynamic systems. Globally, the proposed approach to trust evaluation takes into account categorial knowledge, personal experiences and environmental factors For modeling this in computationally terms, several options are available, ranging from linear, non-recursive functions up to non-linear, recursive models and regression models. In what follows, we discuss three different approaches: Neural Networks, Data Mining and Fuzzy Cognitive Maps.

4.1 Neural Networks

Neural networks (NNs) are largely adopted tools for learning patterns from data, as they provide an effective anticipatory function quantitatively modeling I/O relations. NNs can be used as a *predictor function* on the agent's trustworthiness, thus translated as a reliable DoT expectation. Specifically for categorial trust, a NN is: (*i*) trained on the collection of past experiences, and (*ii*) used for assessing trust of new unknown agents. trust-eval is thus implemented by agents referred as Neural with a 2-layer perceptron. The input of the NN are: the task τ, the environmental conditions E and agent ag_j's manifesta. The output value is the performance prediction for ag_j on τ, namely $DoT_{ag_j,\tau}$.

4.2 Data Mining

Data mining-based categorial trust has been recently proposed by [Burnett *et al.*, 2010] in terms of *stereotypical* trust. The model adopts *stereotypes* as patterns for recognizing

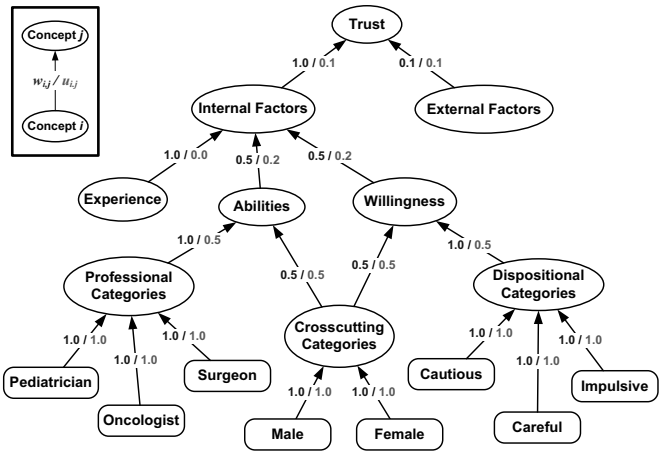


Figure 1: UnFCM implementing socio-cognitive trust

trustworthy agents and associating them an expected DoT. Following this structure, a Stereotrust agent has been implemented and performs categorial reasoning in three phases: (*i*) stereotypes rise from generalization of past experiences and are built using data mining and machine learning techniques; (*ii*) if direct experiences of past interaction with the same trustee are available, then DoT is the average of the previous delegation results; (*iii*) otherwise, given trustee's manifesta and environmental conditions, stereotypes are applied as a filter to determine to which cluster the trustee belongs, thus finding the relative DoT. When a Stereotrust agent has stored an amount of experiences on the same task, it identifies some patterns for recognizing clusters of performers, thus associating them some appraised DoT based on previous delegation results. In Alg. 1, such a process is summarized by the $\mathsf{config}_\tau(\mathcal{E}xp, E)$. Differently from NNs, the data mining approach brings the advantage to generate explicit classification rules. In the concrete implementation, experiences are clustered in 10 bin by equal frequency binning. A decision tree classifier, representing the element ϕ_τ in Alg. 1, is obtained using the C4.5 algorithm[1].

4.3 Fuzzy Cognitive Maps

Fuzzy Cognitive Map (FCM) [Kosko and Burgess, 1998] is a computing technique successfully applied in several domains for modeling knowledge-based systems. FCM is conceived as a graph that models a causal processes by means of concepts and causal relations. The causal impact between two concepts is qualitatively graded by the connection weight ranging in the interval [-1,1]. The value on each concept is iteratively calculated by the weighted sum of the incoming connections, then squeezed by the node activation function f, until convergence is reached on the whole map:

$$A_i(t) = f(\sum_{j=1}^{n} w_{j,i} * A_j(t-1))$$

Differently from NN, FCM exhibits a layout designed by domain experts using an off-line setting. At design time, domain experts identify the relevant concepts of the problem

[1]The WEKA software [WEKA] is adopted for NN and data mining algorithms suite. Further details on the programming model and experimental results are available at `mindraces-bdi.sf.net`.

Algorithm 2 Forward Propagation with Uncertainty

Variables :

Lr : learning rate M : momentum
$Target_{x,p}$: target value of output node x on the pattern p
$Value_{x,p}$: value of node x on the pattern p

procedure $updateWeights(accuracy, epochs)$

1: **repeat**
2: **for ALL** patterns p **do**
3: **for ALL** output nodes j, input nodes i **do**
4: $Error_i(t) = \sqrt{\sum(Target_{j,p} - Value_{j,p})^2}$
5: **end for**
6: **for ALL** the connections from node h to node k **do**
7: $Learn = Lr * Error_h(t) * Value_{h,p}(t)$
8: $\Delta_{h,k}(t) = Learn * u_{h,k} + (M * \Delta_{h,k}(t-1))$
9: $w_{h,k}(t) = w_{h,k}(t-1) + \Delta_{h,k}(t)$
10: **end for**
11: **end for**
12: **until** $Error(t) < accuracy$ **or** $t > epochs$

and quantitatively establish their mutual influences, through weighted connections. Nonetheless, in complex domains, designing an optimal FCM model is troublesome. Both *uncertainty* and *ignorance* may affect the FCM functioning. Experts could not be able to recognize all the mutual influences between concepts, or similarly they could ignore their exact grade. In order to bridge this gap, a learning method for an adaptive, online FCM model has been envisaged.

Uncertainty-driven Learning for FCM

FCM learning is not new and many algorithms have been proposed, based either on supervised and unsupervised learning [Papageorgiou *et al.*, 2006]. We here introduce a novel supervised model, based on the notion of *uncertainty*. We define UnFCM as an extension of the traditional FCM structure aimed at modeling the uncertainty affecting the causal relationship between concepts. We define the *degree of uncertainty*: $u_{h,k} \in$ [-1, 1] as the fuzzy value associated to the weight of the connection $w_{h,k}$ linking the nodes h and k. Alg. 2 shows the learning procedure that exploits such a degree of uncertainty to update the weights of the causal links in the UnFCM. Alg. 2 follows a feed-forward error propagation structure: when the map reaches convergence, the error is calculated as the Euclidean distance of the output nodes from the desirable values on the given pattern (rows 3–5). The error is placed on the input nodes (row 3) and then propagated through the network, according to the combination of weights and uncertainty (rows 6–10).

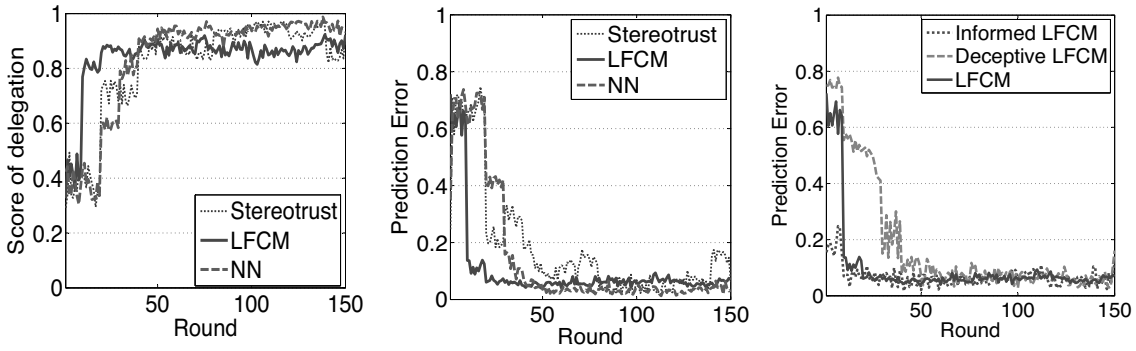
Learning FCM agents (LFCM) utilize the UnFCM illustrated in Figure 1 as cognitive trust module ϕ_τ. This particular structure is shaped on the socio-cognitive model of trust [Castelfranchi and Falcone, 2010], according to which trust rises from a particular configuration of mental attitudes, as expressed by trustors' beliefs and goals related to the trustee behavior. In our setting the nodes adopts an *identity* activation function and *Trust* is the output concept. The two main contributions to trust are characterized as *external* and *internal* factors. Internal factors (i-factors) are related to the internal characterization of the trustee, as believed by the trustor. Two child nodes of i-factors related to the categorial knowledge are considered—thus summing up trustor's beliefs about professional *abilities* and *dispositions*. The categorial nodes refer to particular agent's beliefs as specified by the adopted socio-cognitive model of trust: *Bel(*$Can_{ag_j}(\tau)$*)*, that is trustor belief that ag_j is potentially able to fulfill τ, and *Bel(*$Will_{ag_j}(\tau)$*)*, that is trustor belief that ag_j is potentially willing and persistent in achieving τ. Ability and disposition concepts are then respectively linked to a list of nodes digesting the impact of professional, dispositional and crosscutting categories on the task (see Section 2). Leafs linked to each of the categorial nodes represent the result of perception of trustees' manifesta (Figure 1 shows only three categories for brevity). The i-factor node is also linked to the node *Experience* which is a digest of the personal knowledge of past interactions with the same trustee. External factors (e-factors) summarize the perception of the environmental context E in which each trustee is assumed to operate. This concept is the third constituent belief of the socio-cognitive model, namely *Bel(*$ExtFact_{ag_j}(\tau)$*)* indicating the set of external conditions influencing the task τ to be executed by ag_j.

Figure 1 also shows the weights of the connection (the left side number) which is used by the UnFCM learning. In particular, the weights of the links between the leafs and the categorial nodes assume a pivotal importance for categorial reasoning: they express the potential *impact* of the single categories on the task requirements, thus providing a quantitative measure of how the single categories fit the task execution. Based on the explicit knowledge of both tasks and categories, a function ascribe can be realized as an internal capability of the cognitive trustor to find such a categorial relationship. For instance, having the representation in Table 1, a matchmaking algorithm may find the possible relationship between the categorial profile ⟨*Pediatrician, Careful, Female*⟩, and the task *chickenpox*. In Alg. 1, ascribe is part of the $\mathsf{config}_\tau(\mathcal{E}xp, E)$.

Finally, the degree of uncertainty is defined by design over the levels of the map taking into account the decreasing level of abstraction. Thereby, uncertainty is set to 1.0 for the categorial nodes, 0.5 for professional, dispositional and crosscutting category nodes, 0.2 for ability and willingness nodes and 0.1 for i- and e-factors nodes. The *experience* connection has null uncertainty as we assume direct experience (when available) dominates the categorial reasoning, therefore this link will remain unchanged throughout the learning phases.

5 Experimental Evaluation

The evaluation of the proposed approaches is discussed in this section through experimental analysis. We created a simulated agent society in a medical domain, with 100 trustees randomly selected from a repository of 2500 trustee profiles. The agents profiles have 5 professional, 6 dispositional and 2 crosscutting categories. A set of trustors, with subjective trust evaluation models, interacts with the trustee population over a number of rounds. At each round, trustor's goal is to delegate the assigned task to the best trustee available in the current population. We used *chickenpox* as example of task (see Table 1), for which the best categorial profile is supposed to be ⟨*pediatrician, cautious, female*⟩. The outcome of trustee execution is referred in terms of score, calculated through a matchmaking comparing trustees krypta with task



(a) Score (b) Prediction Error (c) FCM Learning Error

Figure 2: Score chart (a) and error charts (b,c) for different trustors approaches.

	Professional Categories			*Cross Categories*		*Dispositional Categories*		
Agents	Surgeon	Oncologist	Pediatrician	female	male	Cautious	Impulsive	Careful
LFCM	$1,0/.09$	$1.0/-.3$	$1.0/.99$	$1.0/.52$	$1.0/.48$	$1.0/.9$	$1.0/0.0$	$1.0/.58$
Informed LFCM	$.37/.01$	$.42/.01$	$.83/.98$	$1.0/.52$	$1.0/.48$	$.67/.65$	$.33/.0$	$.67/.24$
Deceptive LFCM	$.63/-.06$	$.58/.05$	$.17/.88$	$1.0/.51$	$1.0/.49$	$.33/.64$	$1.0/-.04$	$.33/.21$

Table 2: Initial and final learned weights of the categorial links for different FCM configurations for the task Chickenpox.

requirements (the fulfillment function is omitted).

The *prediction error* for the trustor's evaluation model is defined as the distance of the predicted DoT from the real delegation outcome: $error = |DoT_{ag_j} - score|$. Experiment setting also takes into account the environmental influences, defined as a ρ parameter which indicates the contribution of situated conditions to the final score. Hence, task execution may receive a subjective influence randomly distributed in the range $[-\rho, +\rho]$. An indication of openness is given by the parameter δ, which determines the number of trustees replaced at each round. Finally, a learning interval L sets the number of rounds after which the trustors update their learning model over the new experiences history. In the adopted configuration, we assumed $\rho = \delta = 5$ and $L = 10$. The results are mediated by averaging 20 repeated trials of 150 rounds each.

Delegation Effectiveness

Trustors' approaches are evaluated through the effectiveness in delegating the task to the best performer. Figure 2 (a) shows the trends of the absolute score for Neural, Stereotrust, LFCM agents. It is noticeable a learning phase, that characterizes agents for different intervals. Neural agent learns for ~ 40 rounds, before stabilizing on a mean score of $\sim$.95 (the best among the evaluated agents). Neural finally shows a strong effectiveness in both categorizing agents and in anticipating the delegation result. Neural's prediction error assumes a complementary trend and reaches $< .1$ accuracy after $\sim$32 rounds and resists throughout population changes. Stereotrust agent globally shows comparable performance, although both its score and prediction error never reach a complete stabilization: error trend is irregular and guarantees only $< .2$ accuracy. In fact, the main risk for this agent is that classification stereotypes not always adhere to the population features. Under variable environmental influence, clusters built on the delegation history might be wrongly assorted leading the agent to follow poor delegations. LFCM does not gain the best absolute score ($\sim$.85) but clearly presents the shorter learning phase: it reaches $< .1$ accuracy after 16 rounds, resulting 20% more efficient than Neural[2]. The socio-cognitive structure of the UnFCM integrates the three information levels (personal, categorial and environmental) while learning refines the balance between the information sources. The strength of the LFCM approach is to prune the set of possible trustees according the categorial profile fitting to the task, considerably easing the delegation process.

Learning Categorial Impacts

The second experiment analyzes the effects of the UnFCM initial configuration on the categorial learning. Using the function ascribe (described in Subsection 4.3), we can define an Informed LFCM where the weight of link for a general category *Cat*, with respect to the task τ, is computed as: ascribe(*Cat,*τ). For instance, as indicated in Table 2, the weight of the connection linking the node *Pediatrician* to the node *Professional Categories*, for the task *Chickenpox*, is set to ascribe(*Chickenpox,Pediatrician*) = 0.83. Dually, we define a Deceptive LFCM that uses a misleading distribution of categorial weights. In particular, its weights are set to the complementary value of ascribe. For instance, the connection between *Pediatrician* and *Professional Categories* is set to the weight: $(1 - \text{ascribe}(\textit{Chickenpox,Pediatrician})) = 0.17$. In Figure 2 (c) the prediction errors of these two new models are compared against Informed LFCM.

In Table 2 the initial and the final weights of some categorial links are reported. As a first result, all the three models converge on equivalent configuration of weights showing that the category profile $\langle$*pediatrician, cautious, female*$\rangle$ is dominant for the task *chickenpox*. Differences are noticeable in the learning phases. As expected, Informed LFCM, which

[2]This result was found to be statistically significant by *t*-test, with $p < .05$.

exploit the additional knowledge of the ascribe function, is able to start from $< .2$ accuracy error and its learning phase lasts in only 6 rounds. Conversely, Deceptive LFCM is affected by misleading knowledge and is forced to experience a large number of trustees, spending $\sim$50 rounds in the learning phase. Notwithstanding, it finally minimize the error and attains < 0.1 prediction accuracy.

5.1 Discussion

We pointed out with these examples, the pivotal role of the information sources for reasoning in categorial terms. All the evaluated trust formation approaches were able to perform task delegation based on categorial trust attribution. Categorial evidence with respect to the ongoing tasks *emerge* with different dynamics from each trust model, *without* requiring any initial categorial beliefs. The ability of learning to which extent the single categories fit for a given tasks drastically enhance decision-making and is the base of the emergence of categorial knowledge.

LFCM model attains particular relevance and is worth to be analyzed in more detail. The FCM manages heterogeneous information sources, ranging from personal experiences, to manifesta and external influences, into a single function. In addition, LFCM is able to exploit initial categorial knowledge to boost up its learning features. Thanks to these learning abilities, categories can be revised, or devised from scratch, considering growing personal experiences and the concrete conditions encountered by agents in real domains. Differently from Neural and Stereotrust, a LFCM is further able to maintain the *semantic* of influences between concepts and their connections. This is a pivotal knowledge, and it is explicitly readable from the concepts and the causal links inside the FCM (as in Table 2).

6 Conclusion and Perspectives

As experimental evaluation points out, categorial reasoning, combined with the analysis of experiences and context awareness, provides an effective way to deal with trust formation in dynamic and open systems. Using categories allows agents to abstract form the personal level: delegation is independent of the composition of the population, it is resistant to mutations and replacements, and it also benefits of having reduced categorial information instead of extensive individual experience. As an additional contribute, a comparative analysis of different mechanisms aimed at learning relationships between tasks and categories has been studied, and a novel approach to FCM learning has been presented. Learning allows either to refine decision making, i.e., better anticipating delegation outcomes, either to learn categories form experience. As a remarkable result, system openness can be tackled by agents abstracting from the personal level, with the support of categorial reasoning.

Limitations of the current approach pave the way to future work. To evaluate the scalability of the proposed approach, applications in different domain will be devised. Accordingly, a seamless integration between the deliberative modules and the cognitive mechanisms adopted for trust formation will be be studied at an architectural level.

References

[Bacharach and Gambetta, 2001] Michael Bacharach and Diego Gambetta. Trust as Type Detection. In *Trust and deception in virtual societies*, 2001.

[Barber, 1983] B. Barber. *Logic and the limits of Trust*. Rutgers University Press, 1983.

[Burnett *et al.*, 2010] C. Burnett, T.J. Norman, and K. Sycara. Bootstrapping Trust Evaluations through Stereotypes. In *Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, pages 241–248, 2010.

[Castelfranchi and Falcone, 2010] Cristiano Castelfranchi and Rino Falcone. *Trust Theory. A Socio-Cognitive and Computational Model*. John Wiley & Sons, 2010.

[Fullam and Barber, 2007] Karen K. Fullam and K. Suzanne Barber. Dynamically learning sources of trust information: experience vs. reputation. In *Int. joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-07)*, pages 164:1–164:8, 2007.

[Huynh *et al.*, 2006] 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.

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

[Littman and Stone, 2001] Michael L. Littman and Peter Stone. Leading Best-Response Strategies in Repeated Games. In *IJCAI 2001 Workshop on Economic Agents, Models, and Mechanisms*, 2001.

[Papageorgiou *et al.*, 2006] Elpiniki I. Papageorgiou, C. Stylios, and P.P. Groumpos. Unsupervised learning techniques for fine-tuning fuzzy cognitive map causal links. *Int. J. Hum.-Comput. Stud.*, 64:727–743, 2006.

[Sabater, 2003] J. Sabater. *Trust and reputation for agent societies*. PhD thesis, Univ. Autonoma de Barcelona, 2003.

[Tavakolifard *et al.*, 2008] M. Tavakolifard, S.J. Knapskog, and P. Herrmann. Trust Transferability among similar Contexts. In *ACM symposium on QoS and security for wireless and mobile networks*, pages 91–97, 2008.

[WEKA,] WEKA. `cs.waikato.ac.nz/ml/weka/`.

[Wojcik *et al.*, 2006] M. Wojcik, J. Eloff, and H. Venter. Trust Model Architecture: Defining Prejudice by Learning. In *Trust and Privacy in Digital Business*, LNCS vol. 4083, pages 182–191. Springer, 2006.

[Yu and Singh, 2002] Bin Yu and Munindar P. Singh. An evidential model of distributed reputation management. In *Int. joint conf. on Autonomous agents and multiagent systems (AAMAS-02)*, pages 294–301, 2002.