Reasoning with Categories for Trusting Strangers: a Cognitive Architecture

Matteo Venanzi[1,2], Michele Piunti[1], Rino Falcone[1] and Cristiano Castelfranchi[1]
mv1g10@ecs.soton.ac.uk
{michele.piunti,rino.falcone,cristiano.castelfranchi}@istc.cnr.it

[1] Goal T^3 Group
Institute of Cognitive Science and Technologies (ISTC-CNR) Roma, Italy

[2] IAM Group
Univeristy of Southampton, Southampton, SO17 1BJ, UK

Abstract. A crucial issue for agents in open systems is the ability to filter out information sources in order to build an image of their counterparts, upon which a subjective evaluation of trust as a promoter of interactions can be assessed. While typical solutions discern relevant information sources by relying on previous experiences or reputational images, this work presents an alternative approach based on the cognitive ability to: (*i*) analyze heterogeneous information sources along different dimensions; (*ii*) ascribe qualities to unknown counterparts based on reasoning over abstract classes or categories; and, (*iii*) learn a series of emergent relationships between particular properties observable on other agents and their effective abilities to fulfill tasks. A computational architecture is presented allowing cognitive agents to dynamically assess trust based on a limited set of observable properties, namely explicitly readable signals (*Manifesta*) through which it is possible to infer hidden properties and capabilities (*Krypta*), which finally regulate agents' behavior in concrete work environments. Experimental evaluation discusses the effectiveness of trustor agents adopting different strategies to delegate tasks based on categorization.

1 Introduction

Interaction and *openness* are topics deserving the attention of the research agenda in Multi Agent Systems (MAS): interaction being at the basis of communication, coordination and cooperation, like for instance in virtual societies and networks; openness being at the basis of many of the applicative domains currently developed, like for instance open marketplaces characterized by an ecosystem of mobile devices, services and thousands of exploitable titles and applications. As indicated by many approaches, trust is a pivotal aspect for both interaction and openness. Trust is fundamental for facing the uncertainties typical of open societies, where heterogenous entities are forced to choose whether to interact or not with possibly unknown counterparts. Besides, being at the basis of any interplay, trust is a glue for the whole society: it can promote or prevent interactions of multiple entities, possibly governed by autonomous objectives and capabilities. Even more, trust plays a central role in decision making: it is diriment factor in deciding whether to externalize or not a given activity, or in deciding if a given task can be profitably delegated to another agent.

The downside of trust is that managing it is a costly process for agents. There is a problem of *trust formation*: in order to exploit the benefits of trust, agents need to build a knowledge model able to assess the trustworthiness for each possible counterpart, thus processing additional information about the others. A main issue is in filtering the information sources and in providing a mechanism for evaluating trust on such a basis. Existing literature suggests a couple of alternatives to an agent for assessing trust [7]. The first approach assumes to exploit *personal experience* to analyze how a given agent has performed in past interactions. Otherwise, the shared opinion circulating about a given agent could be exploited in terms of *recommendations/reputation*. In this paper we explore an alternative approach, based on the *reasoning/inference* about the others based on categories of agents. In this direction, we propose categorial trust as a suitable approach to trust formation, and we propose a series of computational mechanisms realizing it in cognitive agents.

Based on a socio-cognitive model of trust [5], we assume that for rationally trusting someone we need a theory of its mind (in case of a cognitive agent) or of its functioning (in case of a more simple artifact). Categorial trust is inspired to an heuristics commonly exploited by humans. It considers the cognitive ability to represent group behavior using general classes or categories of individuals, where categories can be shaped on a specific set of observable features and qualities. The claim of this work is to show that, as in the human case, considering an unknown agent as belonging to a known category allows to infer (or at least attribute) specific internal features for such unknown agent, not directly observable. This means to identify a set of agent's internal features determining how that agent will perform in specific situations. On such a basis, agents may recognize the strict correlation between the internal features of a possible trustee and its pragmatic performances in concrete tasks. In this sense the model recalls the notions of *Krypta* and *Manifesta* [1], according to which manifesta are observable signs for agents' krypta, a sort of internal properties ("qualities", "virtues" or "powers") exploitable to predict/explain their behaviors on specific tasks or activities. Categorial reasoning is provided in order to implement two different level of inference: the former, based on the agentive-personal level, allowing to refine the real capabilities of a given agent based on the analysis of its observable attributes; the latter, based on the societal-categorial level, allowing to refine or create new categories based on the appraised relation between the ability to fulfill a given task and the observable properties belonging to that class of agents. The model proposed in this paper will enable agents to work in both the levels of inference, being part of a cognitive architecture enabling agents to: (*i*) ascribe the effectiveness of a given category for a given task, thus identifying the right trustee on the basis of his potential categorization as expressed by its observable manifesta; (*ii*) assess trust towards a population of unknown agents in dynamic environment conditions, with tasks characterized by changing requirements; (*iii*)assess trust based on partial information about heterogeneous population of agents: a trustor only knows few manifesta for a given trustee.

The rest of the paper is organized as follows. Section 2 surveys related works focusing on the socio-cognitive approach to trust. Section 3 places the research challenge in terms of categorial trust, while Section 4 formalizes a cognitive architecture realizing it and describes a concrete programming model for its implementation. Section 5 presents simulative experiments and results aimed at evaluating the effectiveness of different trust formation strategies. Finally, Section 6 provides final discussion and perspectives.

2 Trusting Agents in Open Systems

Establishing trust in open system requires to effectively build a behavioral model of entities which typically are not known in advance (strangers). From an agent perspective, assessing trust is related to the problem of trust formation, which in open systems refers to the problem to filter a wide spectrum of information distributed within heterogenous sources. Several approaches to trust have been explored in MAS based on experience and reputation [7]. A first strategy relies on the ability to store information of past experiences, and build on such a personal knowledge a subjective model of trust. The same idea has been exploited to assess trust based on statistical analysis [12]. The weakness of these approaches is related the costs in terms of resources needed to explore the whole set of available options before having a direct experience on each available agent. Reputational approaches make use of shared information sources, like certified authorities, reputation and reports. Among others, Sabater at al. proposed a model based on agents' images and reputation [13], according to which social evaluations circulate and are represented as reported evaluations, which are exploited to promote trust formation. Other approaches, as the one explored for instance by [9], makes use of infrastructures making available certified reputation related to each possible trustee agent.

The suggestion to exploit categorial knowledge to assess trust is not new, and it has been theoretically explored for ascertain beforehand the trustworthiness of possible unknown counterparts [2]. In the context of computational models, the work by Wojcik et al. introduced the notion of prejudice filters to perceive particular trustees attributes [14]. 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. allow agents to build stereotypes based on the analysis of past interaction outcomes [4]. Data mining techniques are used to dynamically create classifiers based on personal knowledge. Classifiers are then applied to establish trustworthiness of possible trustees in absence of personal information. As explained in the next sections, the model proposed in this paper revises and extends the use of prejudices and stereotypes in the context of a more general theory of cognitive trust.

The socio-cognitive approach proposed by Castelfranchi and Falcone [5] considers trust as a cognitive process characterized by both relational and graded notions. A pivotal aspect of the socio-cognitive model is that trust formation is a cognitive process based on a series of cognitive ingredients through which the trustor evaluates the trustee in a specific environmental context, by assessing a particular configuration of (positive) expectation and reliance. Trust is a relational notion between a trustor agent (trust giver, ag_i) and a trustee agent (trust receiver, ag_j) which can be established in a given context C, and, most important, about a defined activity or task to be fulfilled (τ):

$$Trust(ag_i, ag_j, C, \tau)$$

. Accordingly, trust is a *graded* construct, and the degree of trust (DoT) comes from the degree of a series of cognitive ingredients, which can be resumed in terms of trustor's beliefs and goals. Summing up, an agent ag_i trusts ag_j about the task τ if DoT overcomes a given threshold σ:

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

Within a group of possible trustees, we assume the trustor will prefer the one having the higher DoT. We omit for simplicity the characterization of trust in terms of additional facts that ag_i has to believe about the trustee and the external conditions (the interested reader can find formalized the approach in concrete implementations, as in [8]). In the particular approach described in this work, such a trustor's beliefs can be assumed as already established once the trustor is able to fill a given trustee in a given category (or class) of agents. Analyzing the wide spectrum of information sources allows ag_i to assess of a series of expectations on ag_j, which in turn makes it possible to assess trust and anticipate its behavior. In this view, trust formation can be assessed on the particular ability of ag_i to analyze a series of ag_j's observable properties (*Manifesta*) and, on such a basis, to infer a theory of ag_j mind (*Krypta*).

3 Cognitive Trust Formation

The approach to cognitive trust proposed in this work assumes two different level of reasoning: the *personal level* which allows to use the information available on the individual trustees, and the *categorial level*, related to the relationship between agents and their categories. Accordingly, for each possible trustee in the system we assume three types of observable information (manifesta). *Professional* and *dispositional* manifesta summarizes internal factors of trust attribution, related in particular to abilities and willingness of a given agent. These features can be exploited at a personal level, i.e., for ascribing a given agent in a specified (professional or dispositional) category. As humans normally do, a particular apparel, particular attitudes or situations can be exploited to find people playing a given role (i.e. a doctor, a dentist, a surgeon) or having a given attitude (i.e. careful, cautious, impulsive). The third class of manifesta considers the information not directly related to professional abilities and willingness, for example being male or female, old or young, religious or atheist, etc. We define this class as "crosscutting" manifesta. In the case of crosscutting manifesta, the relationship with agents krypta has to be *learned* at a categorial level. This is why, for instance, humans form the prejudice that being young, or female, or religious is a better category for fulfilling a series of activities. Summing up, each trustee present in the agent system is assumed as a carrier of three observable properties observable manifesta. For instance a trustee may present features as $\langle$*Surgeon, Cautious, Male* $\rangle$ or $\langle$*Pediatrician, Careful, Female* $\rangle$.

On such a structures, the objective to assess trust is twofold: on the one side it aims to give agents the ability to reason either on the personal level (direct experience), and on the categorial level (categorial experience); on the other side, it aims to show a model of trust built on various levels of information: personal and categorial. We envisage that such an approach may provide an effective heuristic to agents acting in open societies, where the information of prior direct transactions are scarce, and where the possibility to build trust models based on direct experience is infeasible.

In order to design a cognitive model general enough to develop different trust formation strategies, an open scenario has been envisaged. Autonomous agents have to cooperate to carry out a series of tasks inspired to a medical domain, and we assume agents playing two possible roles: patients and medical doctors. At each round, we assume that the *tasks*, inspired by medical diseases, are delegated by patients to doctors. We further assume doctor agents as allowed to enter and exit the system at each time step, thus characterizing the application domain as an *open* system.

3.1 Tasks

The set $\mathcal{T}$ indicates a set of tasks to be fulfilled by patients: $\mathcal{T} = \{\tau_1, \tau_2, ...\tau_N\}$. Each task is characterized by a list of *requirements* needed for its fulfillment: $\tau_j = \langle \tau_{id}, \tau_{Prof}, \tau_{Disp}, \tau_{Cross}, \tau_{State} \rangle$, where $1 \leq j \leq N$ and where requirements are shaped on various dimensions:

- $\tau_{Prof} = \{\alpha_{spec}, \alpha_1, ...\alpha_O\}$ defines abilities (professional) needed to fulfill the task. We assume in particular $\alpha_{spec} \in \tau_{Prof}$ as the pivotal requirement characterizing the task;
- $\tau_{Disp} = \{\omega_1, \omega_2, ...\omega_P\}$ defines willingness (dispositional) to fulfill the task;
- $\tau_{Cross} = \{\kappa_1, \kappa_2, ...\kappa_Q\}$ defines requirements that are not uniquely and immediately related to abilities and dispositions (crosscutting);

Table 1 (a) shows `Chickenpox` and `Appendicitis` as concrete examples of task specification.Task representation includes the structures related to dispositional, professional, and crosscutting categorial requirements. In the `Chickenpox` example, we assume that a specific requirement, called α_{spec}, is the pivotal one to fulfill the task. For instance, to fulfill the `Chickenpox` task, an α_{spec} *pediatr_spec* is needed in order to achieve a result value greater than 0.5. Notice that we assume the cross categorial attribute of being "female" as a task requirement. This

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

Appendicitis	
Abilities	
surgery_spec	99
manual	90
literature	50
technique	90
Dispositions	
availability	90
caution	90
attention	60
Cross	
male	*true*

Male
Crosscutting

Female
Crosscutting

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

Surgeon	
Professional	
surgery_spec:	[99 . . . 100]
manual:	[75 . . . 100]
literature:	[60 . . . 100]
technique:	[60 . . . 100]

Available	
Dispositional	
caution:	[50 . . . 70]
attention:	[50 . . . 70]
availability:	[60 . . . 80]

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

a) Tasks b) Crosscutting cat. c) Professional cat. d) Dispositional cat.

Table 1. Examples of Tasks and Categories specified in a medical domain.

means that, once the task can be fulfilled with a graded result, the contribute of being female consist in an improved outcome, once the fulfillment of a given task ranges from 0 to 100. In concrete implementation, each requirement is modeled as a threshold to be reached by an agent capability in order to be fulfilled[3].

3.2 Categories

$\mathcal{C}at$ are structures indicating a set of abstract categories, or classes, to which agents entering the system may belong. We assume categories as characterized by a list of *features*, shaped on various dimensions and owned by agents belonging to that category.

- Cat_{Prof} indicates professional and pragmatic abilities, grouping together agents specialized in a given activity. For instance, professional categories refers to *Surgeons, Pediatrist, Oncologists*, etc.
- Cat_{Disp} indicates dispositional abilities, grouping together agents characterized by particular attitudes of willingness in fulfilling their activities. For instance, dispositional categories refers to being *Cautious, Careful, Impulsive* etc.
- Cat_{Cross} indicates crosscutting categories not considered in the above mentioned characterization, for instance being *male, female, young, old*, etc.

[3] The choice of task requirements, features and constraints is arbitrary and aimed at showing the functioning and the efficacy of the categorization reasoning, regardless of the compliance of the real medical domain.

Table 1 (b,c,d) shows examples of categories defined in the medical scenario. Professional and dispositional categories include explicit reference to a range of krypta which one may assume for an agent belonging to that category. We assume agents belonging to a given category as having features in the range specified by that category, for instance a *Pediatrician* agent is supposed to have a *manual* ability between 70 and 100, a *pediatr_spec* between 99 and 100, and so on. On the other hands, crosscutting categories only refers to agent's observable manifesta. As said, krypta can not be automatically inferred from crosscutting categories. Hence, the crosscutting manifesta of being *female* initially has an unknown impact on the task fulfillment. The ability to possibly relate the presence of a given crosscutting manifesta to the effectiveness of the agent in fulfilling the task is up to agent reasoning model (it will be described in the next section).

As can be noticed by matching task requirements and category features, each professional category is shaped by design on the requirements of the specific tasks. In particular we assume at least one specializing feature among the professional abilities of a given category related a given task. For instance, we assume the `Pediatrist` category to be related to the `Chickenpox` task by means of the *pediatr_spec* requirement.

4 Agent Cognitive Architecture

We assume an open MAS where the structure $\mathcal{A}g$ indicates a set of agents, each agent possibly entering and leaving the system at any time, and playing the role patient (trustor) or medical doctor (trustee). We assume patient agents are not able to autonomously fulfill the tasks, thereby they need to delegate its concrete fulfillment to a doctor agent. This section provides a formal description of the cognitive architecture through which agents implements trust based delegation.

4.1 Agent Configuration

We assume each agent $ag_i \in \mathcal{A}g$ represented by the following structures:

$$ag = \langle ag_{attr}, ag_{ep}, ag_{goal}, ag_{cog} \rangle$$

where ag_{attr} a list of agent attributes, ag_{ep} represents agent epistemic states (beliefs), ag_{goal} motivational states (goals), and finally ag_{cog} a set of mechanisms realizing cognitive abilities.

Agent Attributes $ag_{attr} = \langle ag_{id}, ag_{role}, ag_{kr}, ag_{mnf} \rangle$ defines a list of attributes owned by agents:

- ag_{id} is the agent identifier (or agent name);
- ag_{role} defines the role actually played by the agent;
- $ag_{kr} = \langle kr_{Ab}, kr_{Will} \rangle$ defines a set of internal properties (*Krypta*), in particular:

- $kr_{Ab} = \{\alpha_1, \alpha_2, ...\alpha_O\}$ defines concrete professional abilities to fulfill tasks;
- $kr_{Will} = \{\omega_1, \omega_2, ...\omega_P\}$ defines concrete dispositional abilities to fulfill tasks;

– $ag_{mnf} = \langle mnf_{Pro}, mnf_{Disp}, mnf_{Cross} \rangle$ defines a list of properties observable by other agents (*Manifesta*), in particular:

- $mnf_{Pro} = \{\phi_1, \phi_2, ...\phi_Q\}$ refers to signals indicating professional abilities;
- $mnf_{Disp} = \{\psi_1, \psi_2, ...\psi_R\}$ refers to signals indicating agent's dispositional attitudes
- $mnf_{Cross} = \{\delta_1, \delta_2, ...\delta_S\}$ refers to signals indicating crosscutting attributes

For instance, professional manifesta may refer to observable signals indicating an agent specialized in pediatrics or in surgery. Dispositional manifesta refers to signals indicating an agent impulsive or cautious. Crosscutting manifesta refers to signals indicating crosscutting categories, i.e., being male or female, etc.

Epistemic States Agent's epistemic states (i.e., beliefs) are represented by the following main structures:

$$ag_{ep} = \langle \mathcal{O}thers, \mathcal{C}at, \mathcal{M}em \rangle$$

$\mathcal{O}thers$ includes an explicit representation for every other agent actually playing inside the system. We assume that an agent ag_i explicitly represent another agent $ag_k \in \mathcal{O}thers$ by storing ag_k's manifesta properties:

$$ag_k = \langle ag_{id}, ag_{mnf} \rangle, \quad ag_k \in \mathcal{O}thers$$

where ag_{id} is the agent identifier, and where ag_{mnf} indicates the signals observed by ag_i upon ag_k.

$\mathcal{C}at = \langle Cat_{Prof}, Cat_{Disp}, Cat_{Cross} \rangle$ indicates the set of categories related respectively to agent professional abilities, dispositions and cross categorial features. In concrete implementation, we assume that the properties observable in a given agent (manifesta) can be automatically retrieved by perceiving the environment. We also assume for the patients a complete knowledge of categories and manifesta in terms of symbolic beliefs.

Finally, $\mathcal{M}em$ builds up the memory of the agent, and it is realized as a belief set storing in patients belief base the results of past delegations.

Motivational States As said, at each round trustor agents (patients) receive a task to fulfill, and for each task they adopt a goal aimed at delegating the activities needed to fulfill it to some trustee (doctors). Such a goal has the following structure:

$$\gamma = \langle \tau, \gamma_{cog}, \gamma_{options} \rangle, \quad \gamma \in ag_{goal}$$

where $\tau \in \mathcal{T}$ is the task associated to that goal, and, from an agent perspective, is given by:

Algorithm 1 Patient delegations process

```
Variables:
    $\tau$ : Task to fulfill.                          $\mathcal{C}at$ : Categories.
    $\mathcal{O}thers$ : Unknown agents.               $\mathcal{M}em$ : Belief set storing results of past delegations.
    $\gamma_{options}$ : Potential trustees.           $task\_cat\_eval$ : Belief set indicating how much a categories fit tasks.

procedure $delegate(\tau)$
    $task\_cat\_eval = \mathsf{ascribe}_\tau(\tau, \mathcal{C}at)$
    $\phi_\tau = \mathsf{fcm}_\tau(\tau)$
    for each $ag_i \in \mathcal{O}thers$ do
        if $\mathsf{matches}_\tau(ag_i, \tau) \neq \perp$ then
            $DoT_{ag_i,\tau} = \mathsf{trust\text{-}eval}(\mathcal{M}em, task\_cat\_eval, \phi_\tau)$
            $\gamma_{options} = \gamma_{options} \cup \langle ag_i, DoT_{ag_i,\tau} \rangle$
        end if
    end for
    $trustee\_agent = findBest(\gamma_{options})$
    send$(trustee\_agent, \mathsf{achieve}, \tau)$
procedure $response(Trustee, \tau, Result)$
    $\mathcal{M}em = \mathcal{M}em \cup \langle Trustee, \tau, Result \rangle$
```

- $\tau_{Prof} = \{\alpha_1, \alpha_2, ...\alpha_O\}$ describes the abilities needed to fulfill the task;
- $\tau_{Disp} = \{\omega_1, \omega_2, ...\omega_P\}$ describes the willingness (dispositions) needed to fulfill the task

Notice that agents ignore τ_{Cross}. In fact, we are assuming a lack of causal knowledge—thus agents which initially are not able to understand how cross categorial features may influence the task. γ_{cog} is the particular cognitive module which is configured to decide to which other agent delegate the task. As will be shown in the next sections, in concrete implementation γ_{cog} is realized through a Fuzzy Cognitive Map (FCM). Finally, $\gamma_{options}$ is a list of possible trustees selected for the delegation. In this case, it represents the options to delegate the task to the trustees. Each element in $\gamma_{options}$ is of the form: $\langle ag_{id}, trust_{id} \rangle$, where ag_{id} indicates a trustee identifier, and $trust_{id}$ represents its related trust value (with $-1 \leq t \leq 1$).

Cognitive Modules In order to find a list of potential trustees for a given task, the trustor has to assess a value of trust each of them. The abstract specification of the trust evaluation model is shown in Alg. 1. It uses a series of cognitive mechanisms and heuristics defined inside ag_{cog}. In particular, ag_{cog} are elements of the type $\langle \Phi, \Psi \rangle$, where Φ represents a decisional module (realized through a Fuzzy Cognitive Map-FCM and described in the next section), and where Ψ includes a set of reasoning abilities, resumed by: (*i*) $\mathsf{ascribe}_\tau$, (*ii*) $\mathsf{matches}_\tau$, (*iii*) fcm_τ, (*iv*) $\mathsf{trust\text{-}eval}_\tau$.
The $\mathsf{ascribe}_\tau$ function, given the specification defined for one task and for each category, allows to quantify the relationship between each category and the specified task:
Definition ($\mathsf{ascribe}_\tau$ - *Associating a Task to Categories*) Let be the representation for a given goal adopted by an agent $\gamma = \langle \tau, \gamma_{cog}, \gamma_{options} \rangle$. Let $\mathcal{C}at \in ag_{ep}$ a belief set indicating professional and dispositional categories. We define: $ascribe_\tau : \mathcal{T} \times \mathcal{C}at \rightarrow ag_{ep}$ as the function $\in \Psi$ finding a series of ex-

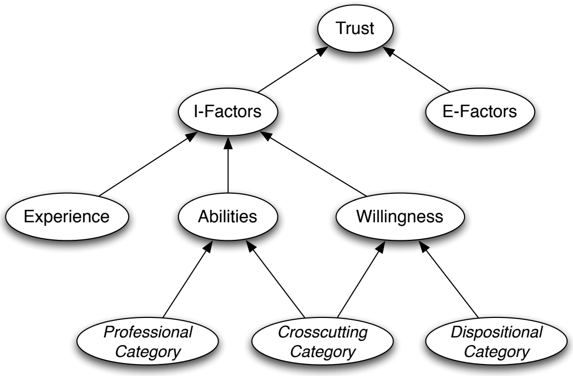


Fig. 1. FCM used by trustor agents to assess the degree of trust of possible trustees.

pressions indicating the matchmaking value between category constraints and the task requirements. In other terms, given the representation of a given task τ, $\mathsf{ascribe}_{\tau,\mathcal{C}at}$ retrieves to which extent the task τ matches the categories $\in \mathcal{C}at$. In concrete implementation, this function produces a set of beliefs to be stored in ag_{ep} relating the task τ to the elements in Cat_{Prof} and Cat_{Disp}. In Alg. 1 (row 1), such a beliefs have the form: `task-cat-eval(Task, Category, ascribe(Task, Category)`.

The $\mathsf{matches}_\tau$ function allows to quantify how a potential trustee belonging to a given category has the required features to fulfill the task or not:

Definition ($\mathsf{matches}_\tau$ - *Matching agent Abilities and task Requirements*) Let $ag_{mnf} = \langle mnf_{Pro}, mnf_{Disp}, mnf_{Cross}\rangle$ the observable properties for an agent $\in \mathcal{O}thers$. Let $\tau \in \mathcal{T}$ a task including a list of agent abilities and dispositions required to fulfill that task. We define: $\mathsf{matches}_\tau : \mathcal{O}thers \times \mathcal{T} \rightarrow \{1, \perp\}$ as the function $\in \Psi$ returning 1 if the categories required for fulfilling the task match the agent properties, $\perp$ elsewhere. In Alg. 1 (row 4), $\mathsf{matches}_\tau(ag_i, \tau)$ is used to verify whether ag_i, according to its manifesta, is matching the requirements needed to fulfill τ.

Given the requirements defined by each $\tau \in \mathcal{T}$, the fcm_τ function allows to configure the appropriate cognitive architecture for that task:

Definition (fcm_τ - *Modulating Architectures for Tasks*) Let the representation for a given goal adopted by the agent $\gamma = \langle \tau, \gamma_{cog}, \gamma_{options}\rangle$. We define: $\mathsf{fcm}_\tau : \mathcal{T} \rightarrow \Phi$ as the function $\in \Psi$ configuring the cognitive map ϕ_τ suitable for evaluating all the possible trustees to which τ could be delegated. In Alg. 1 (row 2), $\mathsf{fcm}(\tau)$ configures a FCM ϕ_τ to be used by the agent to find the best trustee. Given the extent according to which categories match the task τ, and given a cognitive map which is configured with respect to τ, the $\mathsf{trust\text{-}eval}_\tau$ function calculates the trust value for any potential trustee in $\mathcal{A}g$. The output of this function indicates a number resuming the trust value actually assessed for a given trustee.

Definition ($\mathsf{trust\text{-}eval}$ - *Associating trust to a trustee*) Let the representation for a given goal adopted by an agent $\gamma = \langle \tau, \gamma_{cog}, \gamma_{options}\rangle$. Let ag_{ep} the belief base including the set `task_cat_eval`, matching the task τ with the available categories, and the set $\mathcal{M}em$, as the memory of past delegations. Let $\phi_\tau \in \Phi$ the cognitive map configured for the task τ. Then, we define: $\mathsf{trust\text{-}eval}_\tau : \mathcal{O}thers \times \Phi \rightarrow [-1; 1]$ as the function $\in \Psi$ calculating the *trust* value for a given trustee.

In Alg. 1 (row 5), $\mathsf{trust\text{-}eval}(\mathcal{M}em, task\_cat\_eval, \phi_\tau)$ is applied to each possible trustee in $\mathcal{O}thers$ in order to assess its related trust value.

4.2 FCM Trust Attribution

As said, the mechanism underlining trust-eval is realized through a Fuzzy Cognitive Map (FCM) which is configured on the fly by the trustor agent, given the cognitive module $\mathsf{fcm} \in ag_{cog}$ described above. FCMs allow for a flexible computational design of the cognitive model described in Section 2, making it available a straightforward decision making function in different applications and domains [10, 6]. Cognitive maps models a causal process by identifying a series of *concepts* and *causal relations*, being represented as a *weighted graph.* The functioning is governed by Fuzzy Logics [11]: at each computation step, the value of a concept is updated by calculating the impact provided by the other concepts (i.e., the weighted sum of the fuzzy values of the incoming edges). Such a value is squeezed from a specified node's activation function and the computation continues until a convergence is reached.

Fig. 1 shows the FCM used inside the trust-eval mechanism. It is a tree-like structure having *Trust* as root concept. The two main contributions to trust are *external* and *internal factors.* The i-factors are the elements depending on the internal characterization of the trustee, i.e given by trustee's internal capabilities to fulfill the specified task. This node is attached to the two sub-nodes resuming trustee's *abilities* and *willingness.* Each of these nodes is linked to the professional and dispositional categories defined for this domain (see Table 1). The weight of the link reflects the *impact* of the category on the task, as it is computed by the function $\mathsf{ascribe} \in ag_{cog}$.

The adopted FCM uses *identity activation function* and is built so as trust values converge within the interval [-1,1] and no approximation errors is propagated by squeezing the values. We mean the negative subinterval [-1,0] as *mistrust*, namely the case when agent distrusts from delegating the task to another agent. The value 0 means *neutral trust* or absence of trust at all.

This template of the map allows for different types of cognitive evaluations of trust by inactivating or pruning some branches. Indeed, in the special case where also direct experience is considered, a further leaf node *"experience"* is attached to the internal factors. In the scenario discussed in this paper, the trustor uses only *i-factors* branches (related to manifesta and ascribed categories), thus the *e-factors* branches can be excluded from the computation. Instead, *e-factors* branches can be activated for those agents able to understand how the environmental conditions are going to affect the trustee performance.

The concrete implementation of the Alg. 1 is realized as an hybrid architecture. The fuzzy modules through which the cognitive maps are managed is added on top of a BDI engine. The open source project Cog-Trust is used

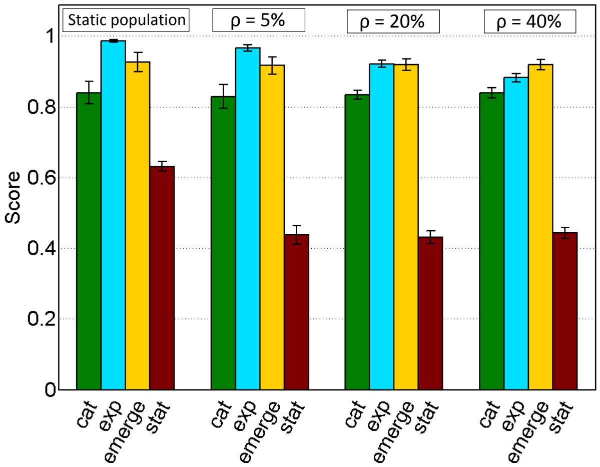


Fig. 2. Mean scores achieved by trustor agents engaged with the task *chickenpox*, in varying conditions.

to implement the cognitive modules, while the BDI engine is realized using the Jason platform [3]. The Jason communication infrastructure is used to realize a simplified contract-net between trustor and trustee agents[4].

5 Experiments

This section presents the experimental evaluation for agents in repeated trials. Experiments observe how different trust formation strategies affect the individual performances of the agents in evolving experimental conditions. Each experiment consists of R *rounds* at the beginning of which, every trustor receives a specific task from the simulator engine. Trustor's goal is to find the best trustee to delegate the task among a population of N possible trustees. An heterogeneous set of trust formation strategies is analyzed. In detail, the following six delegation strategies are considered:

Cat. This strategy is based on the cognitive architecture realizing the categorial reasoning described in Section 4. Categorizing agents are thus able to prune the set of possible trustees looking for those categories that guarantee the best expected outcome. Trust values are computed using a FCM (Fig. 1) including the nodes of internal factors related to abilities and willingness. The map is built according to what said in the previous section and it is populated with the manifesta properties of the trustee retrieved from $\mathcal{M}em$. The FCM mechanism assigns a higher trust value to the trustees who belongs to the professional and dispositional categories better fitting the task requirement. The connections between perceived manifesta and internal FCM nodes are established by the ascribe function, measuring the features matching on the ongoing task.

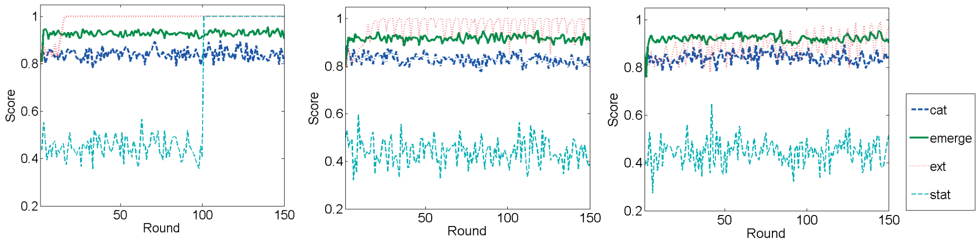
[4] The CogTrust architecture, along with the experiments described in this paper, are available as an open source project at `mindraces-bdi.sf.net`.

Exp. Experience agents add to the FCM used by Cat a further branch summarizing the personal knowledge of the evaluated trustee. Past experiences are resumed for each trustee for the given task. The leaf of the personal experience branch is filled with the values coming from the average of the previous individual performances, as they have been stored in $\mathcal{M}em$.
Stat. The statistic agent uses only personal knowledge. It finds the best trustee on the basis of the history of the previous interactions stored in $\mathcal{M}em$. At each task completion, Stat stores the result value of task fulfillment by the delegated trustee to be used as a index of trust in the next encounters with the same agent.
Emerge. Emerge agents combine categorial and personal reasoning in order to dynamically refine and adjust the trust-eval mechanism used by Cat. Information about crosscutting manifesta is exploited in order to let to *emerge* a set of abstract categories related to the encountered crosscutting manifesta (i.e., being male, female, etc.). Such a crosscutting categories have not a direct relation with abilities and willingness as in the case of professional and dispositional ones, although they concretely influence the performance of the trustee. In order to learn how the emergent category affect trustee's performances, Emerge agents apply a learning mechanisms as part of their trust-eval mechanism. In particular, Emerge agents build clusters inside $\mathcal{M}em$ grouped by crosscutting categories. On such a basis, they try to update the *task_cat_eval* related to the crosscutting categories based on their personal experience.

Fulfillments are measured by absolute scores, referred as the fraction of the highest performance value reachable in the current population for the given task. At the initialization, the simulation engine selects randomly 100 trustees from a repository of 2500 predefined profiles with a random distribution of categories, krypta and manifesta. Openness is measured in terms of population changes. The number of rounds in which the population is fixed forms a *Era*. At the end of each *Era*, $\rho\%$ of the trustee population is replaced by new trustees. In the current setting we use $Era = 5$ rounds. Each experiment is characterized by the score trends averaged for 20 simulations. For simplicity, the experiments have a fixed task (Chickenpox), for which the fulfill function speculates that females perform 10% better than males. Experiments have been run on a machine Inter(R) Core(TM) i5 CPU x64, 2.67 MHz, 6GB RAM, and using Jason 1.3.

5.1 Results

Experiments analyzed how trustor's performance is affected by the frequency and the size of the changes inside the population. We first analyzed agents dealing with a static population and then we progressively increased the ρ parameter to see the effects on the delegation when a small, medium and large part of the population changes. In what follows, we discuss the results for $\rho = 0$, $\rho = 5$, $\rho = 20$ and $\rho = 40$ (Fig. 2).
Fixed Population. Fixed population hypotheses observes trust formation when the population is static (no trustee replacements and $\rho = 0$). In this case direct experiences result a relevant source of information for trust formation. The Exp agent turns to be the best delegator. Being able to exploit the categorization



(a) $\rho = 0$ (b) $\rho = 5$ (c) $\rho = 40$

Fig. 3. Evolution of the trustor scores in any rounds for the task *chickenpox*, varying the ρ parameter, with $Era = 5$ rounds.

reasoning joint to the experience of past delegations, it gets the optimal delegation strategy finding the best trustee within the population (Fig. 2). Stat gets a lower ranking, although its score would be the same of Exp excluding the learning phase spent during the first 100 iterations.

Thanks to the cognitive attribution of trust using categorization and FCM based trust_eval, the exploration of the cognitive agents Cat, Emerge, and Exp is limited to the only specialized trustees (Pediatricians) for the current task. They prune the search space thus wasting less time to find the best performer than the Stat agent. Cognitive attribution of trust based on personal and categorial reasoning allow to quickly stabilize delegation outcomes on the maximum value. The advantage in score of 10% for the Emerge, compared to the Cat agent, is due to the categorial reasoning that let to emerge a preference for females.

Open Populations. Open population hypotheses assume that trustees can leave and can be replaced by others during the simulation. This dramatically increases the probability to face new unknown trustees. Accordingly, openness strongly influences the effectiveness of reasoning on the personal level through direct experiences stored in memory.

When $\rho = 5$, Stat agents show random delegation choices as they are forced to continuously test all the new incoming trustees (Fig. 3(b,c)). The increase of ρ also narrows the gap between Exp agent and the two others categorizer agent: Emerge and Cat. In fact, Fig. 3(b) shows the occurrence of many low scores in the Exp's profile due to the fact that it is not able to further refine the crosscutting categories. $\rho = 20$ is the balance-point, in which Exp and Emerge equalize their scores on 0.93 (Fig. 5, mid-right). For $\rho \geq 40$, Exp finally loses his advantage, as the large replacement of doctor trustees obliges it to compute a new search for the best. Exp totally gets a score of 0.87 while Emerge is the winner with 0.93.

5.2 Discussion

As results point out, agents reasoning on the personal level need to explore the whole population to find the best performer, thus requiring a huge amount of time and resources before reaching an effective result. On the contrary, the combination of categorial reasoning and direct experience promotes an effective exploration strategy. Results confirm that categorial trust is robust to any population change: Cat and Emerge keep the same scores, regardless of the variation of the ρ parameter. The good results of categorizer agents is supported by the computationally efficient implementation of the categorial experience, using the search space $O(|Cat|)$, against $O(|Ag|)$ space required for the individual experience.

Thanks to the FCM structure adopted for trust formation, the distinctive feature of the cognitive trustors is the ability to combine three levels of reasoning: (i) the *categorial level* considers abilities and dispositions of the trustee seen as a member of a known class or category; (ii) the *personal level* is concerned with the direct experiences; (iii) the *environmental/contextual level* which is is concerned with the situation influencing the performances in specific contexts. Facing openness and dynamic populations complicates the delegation, as repeated interactions with the same agent are rare and direct experience mechanisms become increasingly unreliable. This context emphasizes trustor's ability to refine and revise categories, forming *general correlations* and *evaluations* based on the interaction with individuals. Categorization is a twofold reasoning process. Assuming an agent in a class or category is a form of *generalization* from single experiences to form general correlations and evaluations. On the other side, this also allow to transfer, "instantiate", the attributes and features of that general class on a given individual agent.

6 Conclusions

This work describes and evaluates a cognitive architecture based on a model of trust for agents able to reason in terms of categories, against the current approaches which are mostly based on the personal level (reputation, direct experience, observation and statistical analysis). This approach provides an alternative approach to dynamic and open systems. Experimental analysis showed that delegation effectiveness does not depend on the composition of the population, but the model is resistant to mutations and replacements, and it also benefits of efficiency of having reduced categorial information instead of extensive individual experiences.

Limitation of the current approach pave the way to future works. At an architectural level, a seamless integration between the deliberative and cognitive modules will be be studied. The computational model actually forces the developer to specify a FCM template, and then to tune its functioning through an off-line setting of weights and connections. Future work will account the ability of agent to learn connections and adapt the functioning of their cognitive modules on the fly. Another drawback is the need for agents to know a pre-established set of categories ($\mathcal{C}at$). Further studies will explore agents *unifying* personal and categorial level, i.e. autonomously creating new categories from scratch on the basis of individual experiences.

References

1. Michael Bacharach and Diego Gambetta. Trust as Type Detection. In *Trust and deception in virtual societies*, 2001.
2. B. Barber. *Logic and the limits of Trust.* Rutgers University Press, 1983.
3. Rafael H. Bordini, Jomi Fred Hübner, and Michael Wooldrige. *Programming Multi-Agent Systems in AgentSpeak using Jason.* Wiley Series in Agent Technology. John Wiley & Sons, 2007.
4. C. Burnett, T.J. Norman, and K. Sycara. Bootstrapping Trust Evaluations through Stereotypes. In *Int. Conf. on Autonomous Agents and Multiagent Systems (AA-MAS 2010)*, pages 241–248, 2010.
5. Cristiano Castelfranchi and Rino Falcone. *Trust Theory. A Socio-Cognitive and Computational Model.* John Wiley & Sons, 2010.
6. R. Falcone, G. Pezzulo, and C. Castelfranchi. A fuzzy approach to a belief-based trust computation. *Trust, reputation, and security: theories and practice*, pages 55–60, 2003.
7. 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.
8. J.F. H ubner, E. Lorini, L. Vercouter, and A. Herzig. From cognitive trust theories to computational trust. In *Workshop On Trust in Agent Societies (Trust@AAMAS09)*, 2009.
9. 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.
10. B. Kosko. Fuzzy Cognitive Maps. *International Journal of Man-Machine Studies*, 24(1):65–75, 1986.
11. B. Kosko and J.C. Burgess. Neural Networks and Fuzzy Systems. *The Journal of the Acoustical Society of America*, 103:3131, 1998.
12. Michael L. Littman and Peter Stone. Leading Best-Response Strategies in Repeated Games. In *IJCAI 2001 Workshop on Economic Agents, Models, and Mechanisms*, 2001.
13. Jordi Sabater-Mir, Mario Paolucci, and Rosaria Conte. Repage: Reputation and image among limited autonomous partners. *Journal of Artificial Societies and Social Simulation*, 9(2), 2006.
14. M. Wojcik, J. Eloff, and H. Venter. Trust model architecture: Defining prejudice by learning. In *Trust and Privacy in Digital Business*, volume 4083 of *Lecture Notes in Computer Science*, pages 182–191. Springer, 2006.