From Manifesta to Krypta: The Relevance of Categories for Trusting Others

Rino Falcone, Michele Piunti, Matteo Venanzi, Cristiano Castelfranchi
{rino.falcone, michele.piunti, matteo.venanzi, cristiano.castelfranchi}@istc.cnr.it

Institute of Cognitive Sciences and Technologies – ISTC-CNR
Via San Martino della Battaglia, 44 - 00185 - Roma, Italy.

Abstract. In this paper we consider the special abilities needed by agents for assessing trust based on inference and reasoning. We analyze the case in which it is possible to infer trust towards unknown counterparts by reasoning on abstract classes or categories of agents shaped in a concrete application domain. We present a scenario of interacting agents providing a computational model implementing different strategies to assess trust. Assuming a medical domain, categories, including both competencies and dispositions of possible trustees, are exploited to infer trust towards possibly unknown counterparts. The proposed approach for the cognitive assessment of trust relies on agents' abilities to analyze heterogeneous information sources along different dimensions. Trust is inferred based on specific observable properties (Manifesta), namely explicitly readable signals indicating internal features (Krypta) regulating agents' behavior and effectiveness on specific tasks. Simulative experiments evaluate the performance of trusting agents adopting different strategies to delegate tasks to possibly unknown trustees, while experimental results show the relevance of this kind of cognitive ability in the case of open Multi Agent Systems.

Categories and Subject Descriptors: 2.1 [**Agents**]: Multi Agent Systems—*MAS*; 2.1 [**Agents**]: Cognitive Systems; 1.26 [**Social and Information Networks**]: Social and Information Networks; 1.27 [**Social Sciences**]: Multi Agent Based Social Simulation—*MABSS*

Additional Key Words and Phrases: Trust by reasoning, Fuzzy Cognitive Maps, Cognitive Analysis, Open Systems

1. INTRODUCTION

Starting from the Artificial Intelligence field, in particular from the Multi-Agent Systems (MAS) domain, the study of social phenomena like *trust, delegation, adoption*, is now increasing of interest in the more general field of Information and Communication Technologies (ICT). The two main concepts behind this interest are: the autonomy that the new computational entities are developing in a very sophisticated way, and the most recent evolution of the interaction paradigm in computation. Pushing these two concepts of *autonomy* and *interaction* to their logical conclusions, we will have entities pursuing their own goals interacting with others that are at the same way pursuing their goals. In these situations, the primary need for an agent taking part to an intelligent interaction is to assess the trustworthiness of the interacting parts. In addition, we are going towards an interaction scenario in which artificial entities and humans are indistinguishable from each other. In this view, the probability that we have to interact or cooperate with entities we do not have any personal experience with, will be growing, and the ca-

pacity of inferring how trustworthy an agent is, will become a very relevant property of these systems. Many different approaches and models of trust were developed in the last 15 years [Marsh 1994], [Jonker and Treur 1999], [Barber and Kim 2001], [Resnick and Zeckhauser 2002], [Yu and Singh 2003], [Sabater 2003], [Huynh et al. 2006] [Hang et al. 2009], [Ziegler 2009]: they contributed to clarify many aspects and problems about trust and trustworthiness, although many issues still remain to be addressed. The main issue is to understand how trust really works, that is: which are its main sources and basis; how an entity can be considered trustworthy; how the social action of an artificial entity is mediated (in the case of a cognitive agent) by its mental ingredients of trust.

One of the main problems is to analyze the bases of trust: which are the reasons why an agent X has to trust an agent Y? We identify different kinds and nature of basis for trusting other agents:

—*Direct experience* (how Y performed in the past interactions with X);

—*Recommendations* [Yolum and Singh 2003] (other individuals Z reporting their direct experience and evaluation about Y) or *Reputation* (the shared general opinion of others about Y);

—*Inferences/Reasoning* (judgment about Y deriving from a rational reasoning of X not involving direct experience with, recommendation of, or reputation about, the individual agent Y).

In this paper we propose a model inspired by the latter approach, focusing in particular on a reasoning model based on an internal representation of general classes or categories of agents. Categories are defined on the basis of a set of specific constrains in which another yet unknown agent can be inserted. Assuming that an agent can be included in a class or category permits on the one side, to "generalize" from individuals to categories, to form general correlations and evaluations, and, on the other side, to transfer, "instantiate", the attributes and features of that class to a given yet unknown agent. In other words, if there is a way to consider an unknown agent as belonging to a known category (for example there are some signals of that agent reporting/referring an agent's role, profession, but also an agent's attitude, stable disposition, and so on), we can infer (or at least attribute) specific internal features (i.e., not directly observable), that would not be otherwise perceptible, for such unknown agent. We are also considering the fact that there is a strict correlation between agent internal features and agent performances. This does not mean that every agents in the same category will perform exactly in the same way, but, in general, all the agents of the same category/class should ensure a good level of performance about the tasks referred to that category. In this sense we can recall the notions of "Krypta and Manifesta" introduced by [Bacharach and Gambetta 2001], where the so called manifesta of the agents are the signals of their krypta, a sort of internal properties ("qualities", "virtues", "powers") able to predict/explain the agents' behaviors on specific tasks, domains, interactions. These notions do not pertain only to an individual agent: manifesta can be also signs of membership, and thus of qualities that in/for that class of agents those signals mean. In fact, for trusting an agent we need to have a theory of its mind (in case of cognitive agent) or of its functioning (in case of a more simple tool, artifact,

etc.) [Castelfranchi and Falcone 2010]. To do that, we need to identify a set of agent's internal features in order to describe how that agent will perform a task in specific situations. These internal features can be learned by direct experience, but also inferred by the class to which that agent belongs. A careful characterization of these categories and of their relevant features (in particular with respect to the classes of tasks) can lead to predict the performances for agents belonging to a given class when performing a certain task. This is true not only for tasks associated to their class (as generally happens), but also in the case in which agents perform a task relative to a different category. In practice, we have to understand: How can a trustor know if another agent is, for example, a baker, a surgeon or a dentist? Which are the signs (the manifesta) of different categories? How the category define the properties its members (their krypta) belonging to them? And again: how will perform, for example a dentist in a surgeon's task? And will a dentist do better or worse than a baker on that specific task?

This category-based analysis for trust is one of the more diffuse way in which humans rely on and trust other unknown agents in the daily life: we know how to rely on the waiter in a restaurant as waiter, on the driver of the bus as bus driver, on the policeman in the street of a stranger town as policeman, and so on. In this paper we start from the analysis and results of the socio-cognitive model of trust [Castelfranchi and Falcone 1998], [Falcone and Castelfranchi 2002], [Castelfranchi and Falcone 2010]. We extend this model applying it to MAS where agents are able to exploit the knowledge about more or less formalized categories of agents, features and tasks, and where the observable signals of an agent (manifesta) are assumed as clues of its internal properties (krypta). Accordingly, we introduce specific heuristics for:

—Ascribing categories to tasks as a crucial capability for identifying the best trustee on the basis of its potential categorization as expressed by its Manifesta;

—Assessing trust towards a population of trustees in dynamic environment conditions, with different kind of tasks to fulfill;

—Assessing trust based on partial information about a heterogeneous population of agents: trustors only knows few Manifesta for each possible trustee.

The paper is structured as follows: Section 2 summarizes our socio-cognitive approach to trust and extends it with an inference model based on categorization abilities. Section 3 describes the computational model adopted for representing task over categories of agents and Section 4 presents the architecture implemented by cognitive agents able to assess trust based on ascribed categories and situated conditions. Finally, Section 5 evaluates the approach and presents experimental analysis and Section 6 concludes with the discussion of the obtained results and the future developments.

2. CATEGORIES IN THE SOCIO-COGNITIVE MODEL OF TRUST

In this section we summarize the socio cognitive model of trust and describe the extension of this model for reasoning on categorial trust on the basis of a limited set of agent observable features.

2.1 Socio Cognitive Model of Trust

The socio-cognitive model of trust considers trust as a relational construct between the trustor (X), the trustee (Y), about a defined (more or less specialized) task (τ)[1]:

$$Trust(X, Y, \mathcal{E}, \tau, g_X) \tag{1}$$

where are also explicitly present both X's goal (g_X, respect to which trust is activated) and the environment ($\mathcal{E}$) where the relationship is going to take place. $\mathcal{E}$ represents the specific environmental conditions for any involved agents, as they are included in the general set of environmental settings (E). In fact, the trust relationship between X and Y aims at the achievement of the task τ that will satisfy the goal g_X. This achievement can be evaluated by the match between the results coming from the execution of the task (τ), with the goal of the agent X (g_X). In general we have to consider a threshold over which the goal can be considered achieved (this threshold, on its turn, can be dependent from several parameters: the trustor's personality, the relevance of the goal, and so on). The relational construct of trust can be analyzed in terms of the X's mental ingredients of trust that are: the goal g_X, and a set of *main* beliefs:

—$Bel(X\ Can_Y(\tau))$
—$Bel(X\ Will_Y(\tau))$
—$Bel(X\ ExtFact_Y(\tau))$

where:

$\mathbf{Can_Y}(\tau)$ means that Y is potentially able to fulfill τ (in the sense that, under the given conditions, is competent, has the internal powers, skills, know-how, etc); it represents what we call *abilities*.

$\mathbf{Will_Y}(\tau)$) means that, under the given conditions, Y potentially has the attributions for being willing, persistent, available, etc., on fulfilling the task τ; it represents what we call *dispositions*.

$\mathbf{ExtFact_Y}(\tau)$ means that potentially there are a set of external conditions either favoring or hindering Y in realizing the task τ; it represents what we call *opportunities* and, in the specific case of (1), it coincides with the environmental conditions defined as $\mathcal{E}$.

In our model we also consider that trust can be *graded*. In fact, each of the above beliefs can be quantified in terms of "degree of credibility" (about *abilities*, *dispositions*, and *opportunities*). Also for the goal we can consider a value of relevance. We can compose the several grades of credibility in a single degree of trust ($DoT_{X,Y,\tau,\mathcal{E}}$) (see [Castelfranchi and Falcone 2010] for more details). In general, a trust relationship is established when $DoT_{X,Y,\tau,\mathcal{E}}$ overcomes a threshold σ. This

[1]The socio-cognitive model of trust is presented here in its relevant traits. The interested reader can find more in [Castelfranchi and Falcone 1998], [Falcone and Castelfranchi 2002], [Castelfranchi and Falcone 2010].

threshold is also dependent on the value of the goal. So trivially X will trust Y about the task τ if :

$$DoT_{X,Y,\tau,\mathcal{E}} > \sigma \tag{2}$$

Given the previous analysis of the main components of the trust attitude (g_X, $Bel(X\,Can_Y(\tau))$, $Bel(X\,Will_Y(\tau))$, $Bel(X\,ExtFact_Y(\tau))$), we can say that $DoT_{X,Y,\tau,\mathcal{E}}$ is, on its turn, resultant from the several *quantifications* of these components. In what follows we describe a cognitive model allowing trustors to form such relevant beliefs, and in particular to infer the former beliefs (about $Can_Y(\tau)$ and $Will_Y(\tau)$) on the basis of the categorization process.

2.2 Trusting Categories of Agents

Let us now consider a MAS composed by many interacting agents ($ag \in Ag$), each one characterized by a list of own internal features determining agent's behavior in terms of (professional) *abilities* and *dispositions*. We assume that the list of external features of the agent defines its observable state, thus the potential perception of its functional abilities. Similarly to the approach provided in [Bacharach and Gambetta 2001], we define an agent configuration based on the notions of *Krypta* and *Manifesta*:

Definition(*Krypta*). We define krypta as the internal features of an agent, representing agent internal configuration and determining its behavior. We assume that agent's krypta information is not observable by others.

Definition(*Manifesta*). We define manifesta as the external features of an agent, hence as the information which is observable by other agents.

In what follows we assume that manifesta do not determine the agent's behavior in a direct way. Instead, we assume that manifesta are shaped on the internal configuration of the agent and recall its krypta. In other terms, we assume a relation between the agent's krypta and its manifesta: namely, manifesta are the observable signs indicating with a certain approximation internal, unobservable krypta. In doing so, we do not consider the case in which the manifesta are deceptive or wrongly perceived: manifesta are always a hint, a clue of the agent's krypta.

In the described configuration we will consider trustors and trustees as divided groups inside the MAS, thereby if an agent plays the role trustor it cannot plays the role trustee. Agents playing the role of trustor (trust givers) have to identify the best trustee (trust taker) to which the task could be delegated for its fulfillment. We assume that trustors have a partial knowledge of the trustees population, this knowledge is limited to personal experience of past interactions and to the analysis of the available trustee's Manifesta. We also assume that a trustor may assess trust by using its own computational model, i.e. by exploiting statistical information, past experiences, cognitive heuristics, and so on. On such a basis the trustor will delegate the assigned task to the more trustworthy agent and will receive back the value of the trustee's performance as reward. In a population of possible trustees, considering the external conditions $\mathcal{E} \in E$ and the task τ, a trustor X will trust the trustee Y for which the highest assessed degree of trust is assessed:

$$Max_{\tau \in \mathcal{T},(X,Y) \in Ag,\mathcal{E} \in E} \quad \{DoT_{X,Y,\tau,\mathcal{E}}\} \tag{3}$$

Definition(*Tasks*). We define a set of tasks ($\tau \in \mathcal{T}$), each task being identified

by a couple (*action, goal*); where a specific goal (a state of the world to achieve or to maintain) can be reached by that specific (simple or complex) *action*.

Theoretically, the tasks are defined by a set of actions' *requirements* identifying those agents' *internal features* useful for successfully performing the actions—thus achieving the specified *goal*. These *requirements* are referred to both the *professional* and *dispositional features* of the agents. In practice, on the basis of the actions' *requirements* of the tasks, are directly individuated the abilities and dispositions of the agents (or of classes/categories of them) needed to successfully realize those actions in order to achieve the specified goal. We also assume that *external* (environmental) conditions in which the task is realized could affect the performance of the delegated trustee.

Definition(*External Factors*) We define the external factors ($\mathcal{E} \in E$) as those contextual conditions determining the situation in which the task is executed.

In fact, we assume two potential influences of the external environment on the trustees' performances: on the one hand, different environmental conditions could lead to different results in the world, even starting from the same trustee's actions (it depends on the different composition of the trustee's actions with the environmental conditions); on the other hand, different environmental conditions could change the trustee's actions themselves. In this case it results different the final trustee's actions, rather than their composition with the environment. In the MAS described above, the problem to infer trust towards possible trustees is in anticipating their performances, that is to evaluate a trust value for a trustee only showing a list of its Manifesta. Such a trust value can be used as reliable indicator of the trustee's performance on a given task. An available option for a trustor to do this is to ascribe a category to any possible trustee and, on such a basis, build a theory of that agent. We assume categories characterizing classes of agents in behavioral terms.

Definition(*Categories*) We define $\mathcal{C}at$ as the set of categories, each category being determined by a set of features's *constrains*.

Members of a given category have their profile's features bounded in a certain interval. In this characterization, a trustor can include a trustee in a given category by exploiting its manifesta. In doing so, a trustor can assume that the trustee has a range of features which is proper for the specific task requirements. Thereby, given a category, an agent may anticipate to some extent which performance the agent belonging to that category is going to realize.

We remark that, inside each category, the agent performances may belong to a range of values, varying from low to high effectiveness due to the actual grade of agents' features. By knowing the category to which an agent belongs, a trustor knows just approximately the internal features of the agent: such a knowledge is given by the range of values (constrains) characterizing that category. Indeed, the actual skills and behaviors of the trustees are determined by a set of their internal *features* which, by definition, are not observable by others (Krypta). It is worth to remark that even ascribing a category to an agent on the basis of its manifesta, a trustor continues to ignore the real values of the internal *features* of the trustee (that is, agent's manifesta can just refer an approximate value of its krypta).

On these basis, we envisaged a cognitive architecture enabling agents to trust

Dentist		
Features	Low	High
manuality:	[40 . . . 60]	[80 . . . 100]
dentistry_spec:	[99 . . . 100]	[99 . . . 100]
expertise:	[40 . . . 90]	[80 . . . 100]
problem solving:	[60 . . . 80]	[80 . . . 100]

Surgeon		
Features	Low	High
manuality:	[70 . . . 80]	[85 . . . 100]
surgery_spec:	[99 . . . 100]	[99 . . . 100]
expertise:	[60 . . . 80]	[80 . . . 100]
problem solving:	[60 . . . 80]	[90 . . . 100]

Otorhinolaryngologist		
Features	Low	High
manuality:	[40 . . . 60]	[70 . . . 100]
ent_spec:	[99 . . . 100]	[99 . . . 100]
expertise:	[60 . . . 80]	[90 . . . 100]
problem solving:	[50 . . . 70]	[80 . . . 100]

Oncologist		
Features	Low	High
manuality:	[50 . . . 70]	[75 . . . 100]
oncology_spec:	[99 . . . 100]	[99 . . . 100]
expertise:	[40 . . . 60]	[90 . . . 100]
problem solving:	[50 . . . 70]	[90 . . . 100]

Radiotherapist		
Features	Low	High
manuality:	[40 . . . 60]	[80 . . . 100]
ent_spec:	[99 . . . 100]	[99 . . . 100]
expertise:	[50 . . . 70]	[80 . . . 100]
problem solving:	[40 . . . 60]	[90 . . . 100]

Table I. Professional categories

through categories as built on the following main functions:

—$\mathsf{Ascribe}_{\tau,cat}$. Given the description of the current task τ, and the category cat, this function calculates the degree of the match between the constrains of cat and the requirements of τ (see Subsection 4.1).

—$\mathsf{Matches}_{\tau,cat,ag}$. Given the categories ascribed for each task τ, and given the list of manifesta owned by an agent (ag), this function allows to verify whether the trustee has the required internal features to fulfill the task or not (see Subsection 4.2)

—$\mathsf{TrustEval}_{\tau,ag,Bel,\mathcal{E}}$. Given a task τ, an environmental influence $\mathcal{E}$ and an agent (ag), and given the trustor belief base (Bel) storing the history of past interactions, this function allows the trustor to synthesize a trust value for ag. In concrete implementations, trust evaluation for the cognitive trustors is realized through a mechanism based on Fuzzy Cognitive Maps (see Subsection 4.3).

The details of these functions will be matter of Section 4. It has to be remarked that this approach allows agents to reason in a twofold level, namely in a categorial and in a personal level: the former, ascribing categories to agents based on their manifesta, the latter, including the belief base of the agent to exploit the information about past personal experience. As said, the described cognitive heuristic also allows to evaluate external conditions and their influence on agents' performances.

3. A COMPUTATIONAL MODEL FOR TRUST BASED ON TASKS AND CATEGORIES

The cognitive approach to trust assessment previously described allows trustors to combine different information sources. The task is ascribed to a list of suitable categories which drive the selection of possible trustees. In doing so, the cognitive trustor analyzes trustees' manifesta, on the bases of these, extracts some measures

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

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

Available	
caution:	[50 . . . 70]
attention:	[50 . . . 70]
availability:	[60 . . . 90]

Impulsive	
caution:	[30 . . . 50]
attention:	[40 . . . 60]
availability:	[60 . . . 80]

Distracted	
caution:	[40 . . . 60]
attention:	[20 . . . 40]
availability:	[50 . . . 70]

Reluctant	
caution:	[60 . . . 80]
attention:	[60 . . . 80]
availability:	[30 . . . 50]

Table II. Dispositional categories

Dental Operation	
Abilities	
manuality:	80
dentistry_spec:	99
expertise:	80
problem solving:	80
Dispositions	
availability:	90
caution:	90
attention:	90

Appendicitis	
Abilities	
manuality:	100
surgery_spec:	99
expertise:	50
problem solving:	100
Dispositions	
availability:	70
caution:	90
attention:	60

Otitis	
Abilities	
manuality:	80
ent_spec:	99
expertise:	100
problem solving:	100
Dispositions	
availability:	70
caution:	90
attention:	90

Table III. Example of tasks, each indicating a list of professional and dispositional requirements.

of their professional and dispositional capabilities, also taking into account possible environmental influences over the task execution. Before providing the details of the mechanisms at the basis of the cognitive architecture, in what follows we first describe the structures used by agents for reasoning in terms of categories and tasks.

3.1 Categories

The set of categories $\mathcal{C}at$ models a shared explicit information inside the MAS. We consider *professional* and *dispositional categories*: They define respectively the common *abilities* (or capabilities, skills) and *dispositions* (or personality traits, willingness, intentional attitudes) of their belonging agents.

Each category is defined by a set of *features' constrains*, where each *constrain* bounds a certain agent *feature* to range within a minimum and a maximum value—being bounds defined by the interval [0, 100]. To be fully comparable, these categories are designed on the same set of features; for example the features of the *professional categories* are: {*manuality, specialization, expertise, problem solving*}, where specialization refers to a *specializing* feature. Similarly, the *dispositional categories* are specified by: {*caution, attention, availability*}.

Professional categories are referred to the medical domain and are reported in Tab. I. Categorial features, requirement intervals and constraints are not referred to experimental data, but they are inspired to a general common way of reasoning. This choice is aimed at showing the functioning and the efficacy of the categorization reasoning, regardless to the compliance of the real medical domain. As it will be discussed in the model evaluation (Section 5), arbitrariness would compromise the results of our model only partially[2].

[2]It is clear that there will be various dimensions for trustworthiness and that there will be various

As said, any professional category is characterized by a *specializing* feature, that is the feature *professionalizing* the category. For instance, in the *Otorhinolaryngologist* category, *ent_spec* refers to otorhinolaryngology specialization[3].; the *Surgeon* category is characterized by *surgery_spec*, and so on (see Tab. I). On the basis of the professionalizing feature, a taxonomy of categories exists. Each category is indeed divided into two subcategories—*Low* and *High* meaning classes of agents with lower or higher skills for the same specialization. This allows to better observe how categorizer agent addresses trust to the most professionally specialized category for the given task.

Dispositional categories model a particular character profile of an agent. We consider six dispositional categories, as reported in Tab. II. As for the professional categories, also dispositional categories are designed on a basic set of features, that are: {*caution, attention, availability*}. Notice that each dispositional category is designed to implicitly *promote* a specific feature (e.g., category: *Cautious* $\rightarrow$ feature: *caution*) while the dual category *penalizes* that feature (e.g., category: *Impulsive* $\not\rightarrow$ feature: *caution*). This allows to better highlight the relation between agent's dispositions and task requirements in terms of behavioral attitudes. Internal dependencies among features are considered too. For example, the category *cautious* has a high value for the feature *caution* and a lower value for the feature *availability* (as we assume that the cautious agent will be lower in performing a task at the expense of his availability).

Having an agent member of a high professional category does not necessarily mean that it will perform better than any members of the respective low professional category. This is due to the fact that, as identified in the socio-cognitive model of trust, we assume each agent belonging to exactly two categories, professional and dispositional ones. Thus, the evaluation of agent's performance on a task depends not only on his abilities (*features* referred to the *professional* category) but also on his dispositional attitudes (*features* referred to the *dispositional* category) in addition to the potential positive or negative influence of the environment. Maybe an agent presenting very high professional features could offer a very low dispositional attitudes.

3.2 Tasks

Tasks are automatically provided to the agents by a system engine. A task is represented in the agent's knowledge as a set of *requirements* (both *professional* and *dispositional*) that should be satisfied by the performer's features for successfully realizing that task. A *threshold value* is associated to each of these features. The threshold value is the minimum value (for that feature) the trustee must supply in order to satisfy that specific feature.

The tasks considered in our medical scenario are the ones described in Tab. III. As already said, task specification can include a *specialization* requirement which specializes the task over a specific professional category. For instance, the *otitis* task is characterized by an *ent_spec* feature, which refers to otorhinolaryngology

values on these dimensions, with individual differences. The interesting thing to analyze is: how to compute them and which will be the emergent result of these different values?

[3]We use *ENT* (Ear, Nose and Throat) as synonym of otorhinolaryngologist for simplicity.

Algorithm 1 FulfillTask function

Variables:

τ : task $\in \mathcal{T}$
ag_i : agent $\in Ag$
Kr_i : set of krypta features of Ag_i
R_τ : set of requirements of task τ
$V[size(\mathcal{T})]$: array to store the match values between the features of R_τ and Kr_i
$missing$: number of requirements of R_τ missing in Kr_i
$covered$: number of requirements of R_τ that are also in Kr_i
Returns : the average match value of the requirements of R_τ against the krypta features of ag_i contained in the Kr_i. This value is scaled on the number of missing features

function FulfillTask(τ, Ag_i)

1: $F_i = getFeaturesSet(Ag_i)$
2: $R_\tau = getRequirementsSet(\tau)$
3: $missing = 0$, $total\_overlap = 0$, $covered = 0$
4: **for** all $r_k \in R_\tau$ **do**
5: **if** $r_k \notin Kr_i$ **then**
6: $missing = missing + 1$
7: $V[k] = 0$
8: **else**
9: $f_i = getFeature(r_k, Kr_i)$ /*feature of Kr_i corresponding to r_k*/
10: $V[k] =$ ***featureMatching***(f_i, r_k)
11: $covered = covered + 1$
12: **end if**
13: **end for**
14: $mean = \frac{\sum_{v_k \in V}(v_k)}{covered}$
15: **return** $result = mean * \frac{1}{1+missing}$ /*scaled by missing features*/

Variables:

r : requirement
f : feature
Returns : the matching value of f against r

function $featureMatching(f, r)$

1: **if** $value(f) > value(r)$ **then**
2: $result = 1$ /*requirement satisfied*/
3: **else**
4: $result = \frac{value(r)}{value(f)} * 100$ /*requirement partially satisfied*/
5: **end if**
6: **return** $result$

specialization present in the *Otorhinolaryngologist* category. In this configuration, tasks and categories can be mapped each other on the basis of their specializing properties.

3.3 Trustee Agents and Task Fulfillment

Tasks can be executed only by agents playing the trustee role. A task is accomplished by a trustee with an action performed over the artifacts (Ar) representing the MAS environment. We define the value resulting from the task performance as a *score* computed by the actual features of the performer, that is by its krypta. In fact, a task is potentially well fulfilled only when the all the thresholds of its requirements are exceeded by the corresponding trustee's features.

We use the FulfillTask function, showed in Algorithm 1, to quantify the value of fulfillment (score). This function is stateless, and it is implemented inside the environment artifacts (Ar), through which the actions are concretely executed and task achieved. The function provides a numerical score proportional to the matching

value between task's requirements R_τ and trustee's features F_i. R_τ and F_i include both abilities and dispositions (Algorithm 1, row 1,2). A second function (*featureMatching*) is then used to provide the concrete matchmaking value between agent's features and task requirements (Algorithm 1, row 10). Different techniques could be specified to define the matchmaking: *featureMatching* in Algorithm 1 utilizes a simple comparison that quantifies the overlap between each task requirement and the related agent's features. Finally, the fulfillment value (*score*) is the sum of all the single overlaps between agent's features and task's requirements, normalized to 100 and scaled over the number of missing features (row 14,15).

4. AGENT COGNITIVE MODEL

In this section some of the relevant aspects characterizing the architecture of the cognitive agents are presented. In particular, the computational model implementing the functions defined in Subsection 2.2 is described.

4.1 Ascribing Categories to Tasks

$\mathsf{Ascribe}_{\tau,cat}$ is the cognitive function used by the trustors for comparing the requirements of the task τ with the constrains characterizing the category *cat*. The ascribe mechanism is showed in Algorithm 2. First, the category constrains and the task requirements are retrieved using their representations (rows 1,2). Then, every task requirement is compared with the constrains of *cat*. For each requirement r_k a matching value is calculated (rows 4-11). If the task requirement meets in the category constrains, a matchmaking value is calculated using the subfunction *constrainMatching* (row 9). Otherwise the requirement r_k is considered as *not* satisfied, and the variable *missing* is incremented. As for the function FulfillTask, different techniques could be specified to define the matchmaking between category's constraints and task's requirements: *constrainMatching* in Algorithm 2 utilizes a simple comparison to quantify the overlap between each task requirement and the related agent feature. The sum of all the partial overlaps is then normalized to 100 and scaled over the number of missing requirements (rows 13,14).

4.2 Associating Agents to Tasks trough their Categories

Given a task descriptor τ, and given the manifesta exhibited by a trustee *ag*, the $\mathsf{Matches}_{\tau,cat,ag}$ function defines whether or not the categories to which *ag* belongs are suitable to fulfill the task τ. In doing so, this function checks whether the total of all the contributes provided by the categories owned by the agent *ag* allows a trustor to trust *ag*. The categories of the agent are derived from the its manifesta (row 3), and every single value ascribing the category to the task is used to increment a global *match* value (rows 6-8). Finally, the function returns *true* if this match value does overcome a given threshold σ (row 9).

Notice that Matches satisfies the general statement of Equation 2: the *match* value represents in this case an approximation of *DoT* computed on a categorial level. For simplicity, we just outline here the sub-function *findCategory*: concretely, given trustee's manifesta ($mnf_k \in M$), this function retrieves the category to which the agent belongs using a set of predefined rules. In the adopted configuration, *findCategory* does not introduce further uncertainty, categories are directly mapped to manifesta and thus associated to agents with a rate 1:1. That is, in the concrete

Algorithm 2 Ascribe function

Variables:

τ : task $\in \mathcal{T}$
cat : category $\in \mathcal{C}at$
C : set of constrains of cat
R_τ : set of requirements of task τ
$V[size(\mathcal{T})]$: array to store the match values for each requirement of τ evaluated on cat.
Returns : the average match value of the requirements of R_τ against the constrains of C, scaled on the number of missing constraints

function Ascribe(cat, τ)

1: $C = getConstrainsSet(cat)$
2: $R_\tau = getRequirementsSet(\tau)$
3: **for** all $r_k \in R_\tau$ **do**
4: **if** $r_k \notin C$ **then**
5: $V[k] = 0$
6: $missing = missing + 1$
7: **else**
8: $c_k = getConstrain(r_k, C)$
9: $V[k] =$***constrainMatching***(r_k, c_k)
10: $covered = covered + 1$
11: **end if**
12: **end for**
13: $mean = \frac{\sum_{v_k \in V}(v_k)}{covered}$
14: **return** $result = mean * \frac{1}{1+missing}$ /*scaled by the missing constraints*/

Variables:

r : requirement
c : constraint
Returns : the matching value of r against c

function $constrainMatching(r, c)$

1: **if** $value(r) > upperBound(c)$ **then**
2: $result = 0$ /* requirement not satisfied*/
3: **else**
4: **if** $value(r) < lowerBound(c)$ **then**
5: $result = 1$ /* requirement satisfied*/
6: **else**
7: $\Delta = upperBound(c) - lowerBound(c)$
8: $result = 1 - \frac{value(r) - lowerBound(c)}{\Delta}$
9: **end if**
10: **end if**
11: **return** $result$

implementation of the MAS described in the next section, it is straightforward to retrieve categories from the corresponding manifesta. The possibility to have an uncertain attribution of categories is deemed for future work. It is worth to remark that this function allows to focus on the *discriminant* requirements, thus narrowing the delegation search space only to those trustees exhibiting just the proper professional categories and avoiding inappropriate delegations to unsuitable trustee [Castelfranchi and Falcone 1997].

4.3 Assessing Trustworthiness through Fuzzy Cognitive Maps

In order to compute trustworthiness of trustees belonging to Ag, cognitive trustors adopt the function $\mathsf{TrustEval}_{\tau,ag,Bel,\mathcal{E}}$. For each trustee filtered by the function Matches, this function combines the information inferred on the categories owned by the trustee with the situated environmental influences ($\mathcal{E} \in E$). The function realizes the computation of $DoT_{X,Y,\tau,\mathcal{E}}$, which is finally ranked to find the best

Algorithm 3 Matches function

Variables:

ag : agent $\in Ag$
τ : task $\in \mathcal{T}$
σ : threshold
M_{ag} : set of ag's manifesta
C_{ag} : subset of $\mathcal{C}at$ of the categories to which ag belongs, retrieved by the ag's manifesta
$match$: sum of the values ascribed to τ for each category in C_{ag}
Returns : true if $match$ overcomes σ, *false* otherwise

function Matches(ag, τ, σ)

1: $M_{ag} = getManifestaSet(ag)$
2: **for** each $mnf_k \in M_{ag}$ **do**
3: $C_{ag} = C_{ag} \cup$ ***findCategory***(mnf_k)
4: **end for**
5: $match = 0$
6: **for** each $cat_k \in C_{ag}$ **do**
7: $match = match +$ Ascribe(cat_k, τ)
8: **end for**
9: **return** $match \geq \sigma$

trustee, as showed in Equation 3.

TrustEval is assumed to merge all the contributions to trust, as they are identified in Section 2. For doing this, several options are available, ranging from linear, non-recursive functions up to non-linear, recursive mechanisms as Neural Networks (NN). The architecture described here adopts the non-linear mechanism of Fuzzy Cognitive Maps (FCM) [Kosko 1986]. The main advantage of using FCMs is to be a structure that offers a flexible computational design of the cognitive trust model, as well as it is suitable for different applications and domains. FCM is indeed a cognitive map further enriched with Fuzzy Logics [Kosko and Burgess 1998]. In general, cognitive maps model a causal process by identifying *concepts* and the *causal relations*, represented as a *weighted graph*. The causal effects can be determined by domain experts at design time by simply weighting the links. In this case the FCM has the layout shown in Fig. 1, where the designer has defined the impacts for the internal factors given by the factor 1.0 for *Experience* and *Abilities*, and by the factor 0.5 for the *Dispositions*.

Benefit of FCMs are also their robustness and adaptability. In this case, the map is designed as a tree-like graph with *Trust* as root concept (see Fig. 1). Following fuzzy reasoning rules, at each computation step the value of any concept (node) is updated by calculating the impact provided by the other concepts (i.e., the weighted sum of the incoming edges). Similarly to a NN, such a value is then squeezed using the node's activation function, thereby introducing non-linearity. The computation continues until the convergence is reached, that is until the updates do not significantly change the node values anymore. The use of particular FCM configurations allows to flexibly adopt different strategies of reasoning. Indeed, by inactivating or pruning some branches of the map, different kinds of trust evaluation can be straightforwardly performed. For instance, in the case of the simple cognitive agent which uses only categorial reasoning through manifesta information, the branch *external factors* is excluded from the computation. Dually, agents using the personal level, based on the direct experience only, may refer to the *experience* node, thus cutting off the categorial branches and the ones related to the external factors.

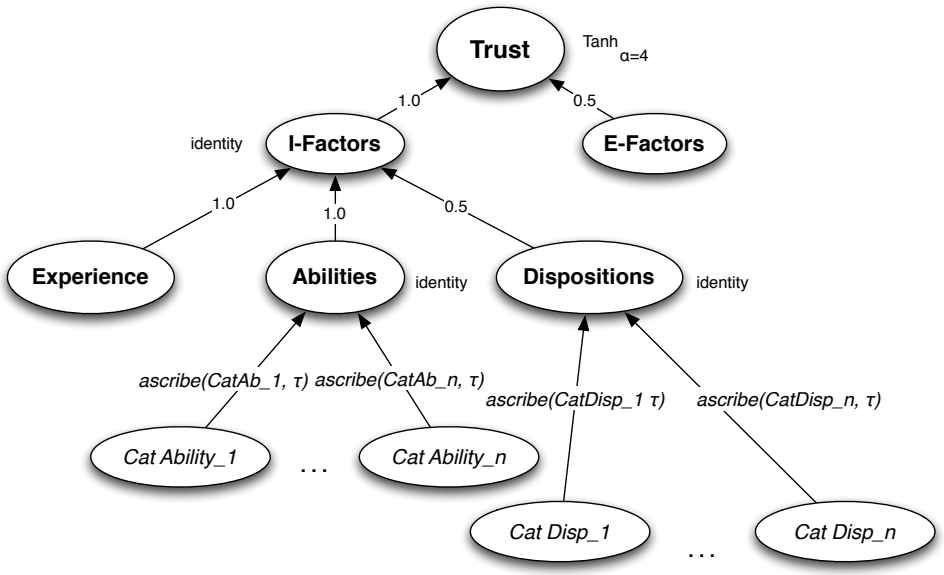


Fig. 1. **Fuzzy Cognitive Map.** Configuration used by the cognitive trustors to assess trustworthiness of trustees.

A further important property of the FCM is to maintain a unambiguous semantic for the involved structures. Differently from NN, the configuration of a FCM can be read, understood and quantified at execution time in terms of concepts and concrete influences between concepts. FCM can be also used in conjunction with machine learning techniques inspired either to unsupervised or supervised algorithms used by NN [Papageorgiou et al. 2006]. Learning is aimed at dynamically refine the weights of the causal relationships between concepts, introducing recursive loops between input and output nodes. FCM with learning features provide a fully adaptable decision making module. Our current works are showing that learning FCM techniques allow to better adhere to the problem domain, for instance widely improving the accuracy of the model in assessing trust values which finely reflect the numerical anticipation of the delegation results.

In Fig. 1, the map points out the two main contributions to trust that are *external* and *internal* factors (realized by the E-Factors and I-Factors nodes). By design choice, these factors affect the trust node with a fixed weight of 0.5 and 1.0 respectively. The map does not make use of feedback loops for learning, nor consider circular influences between concepts[4]. Internal factors are all those elements depending on the internal characterization of the trustee, as believed by the trustor. The I-Factors node is linked to the node *experience* representing the knowledge of past direct interactions with the same trustee. I-Factors also has two further child

[4]It has to be considered that there are cases in which a circular relation between external factors and abilities exists. For instance, *driving* abilities are influenced in presence of *ice on the road*, *work overload* may influences the ability to *be effective*, and so on. However these cases are not considered in our FCM model at the moment.

nodes related to the specific categories considered in the trustor's domain—thus summing up trustor's beliefs about professional *abilities* and *dispositions*. They refer to particular agent's beliefs, which have been identified as $Bel(X\ Can_Y(\tau))$, $Bel(X\ Will_Y(\tau))$ in Subsection 2.1. Each of these two concepts is then respectively linked to a list of nodes indicating the professional and dispositional categories in $\mathcal{C}at$ (see Tab. I and Tab. II). The weight of the links reflects the *impact* of the category on the task requirements and is computed by the function Ascribe.

E-Factors node summarize trustor's perceptions about the context conditions in which each trustee is assumed to execute the actions to fulfill the task. In concrete implementation of the system, we assume that the influence of these conditions on the performance can be positive, negative or null. We also assume the trustor knowing the external factors $\mathcal{E} \in E$, for each available trustee and with no uncertainty. We also assume this contribute being stored in the trustor's belief base in terms of $Bel(X\ ExtFact_Y(\tau))$, as identified in in Subsection 2.1. In the special case where also direct experience (of the trustor with the same trustee) is considered, a further leaf node *experience* is linked to the internal factors.

The convergence on the root node is the output of the FCM and provides trustworthiness value for a given trustee. As trust values range within the interval [-1, 1], a good configuration for the FCM is to adopt an *hyperbolic tangent activation function* [Bueno and Salmeron 2009]. In our architecture, the root node uses an *hyperbolic tangent* with scale factor $\alpha = 4$, while all the internal nodes are provided with an *identity activation function*. By doing so, we do not loose dependencies between internal nodes and furthermore we make sure that no approximation error is computed and thus propagated by squeezing the values during the convergence process. As for trust values, we mean the negative subinterval [-1, 0] as *mistrust*, namely the case when agent distrusts to delegate the task to the trustee. The middle value 0 means *neutral* trust or *absence* of trust. Neutral trust may be possible either due to lack of information (leaves set to 0) or to divergent sources —namely, positive evaluation about professional categories opposite to negative evaluation of dispositional categories and/or environmental influences.

5. EXPERIMENTS

The evaluation of our approach has been conducted through experiments while measuring the performance of our socio-cognitive model against alternative statistical model commonly used for evaluating trust and reputation. This section describes the experimental setting, presents the obtained results and discusses their significance.

5.1 Experimental Setting

The scenario is designed as a time-stepped simulation in which participant agents playing the role of trustor and trustee have to cooperate to carry out a number of tasks in the medical domain. At the beginning of each round every trustor receives a task ($\tau \in \mathcal{T}$) from the simulation engine and it has the goal to achieve the highest payoff from performing the assigned task. Furthermore, trustors are assumed they are not able to autonomously fulfill the task but they need to find the best trustee to which delegate the task execution. Experiments have been conducted using a population of 100 trustee agents. In the first set of experiments we consider three

tasks: *dental operation, appendicitis, otitis*, as defined in Tab. III. We assume that trustees can accept more than one delegation coming from different trustors at every round. In addition, we make sure that at least 30% of the available trustees is guaranteed to belong to the professional category specialized for the task. For instance, when the task is *otitis*, there are at least 30 trustees belonging to the categories specialized with $ent\_spec \geq 99$. The course of the experiments has been fixed to guarantee the agents to stabilize their scores, thus experiment length has been fixed in 200 rounds. The parameter δ measures the *influence* of the environment on the task performance. This influence can vary between $-\delta\%$ and $+\delta\%$ in the final score. A trustee (ag_i), associated to the context (ϵ_i), in fulfilling the task (τ) is thus affected by positive, negative or irrelevant context conditions, which influence the performance for a rate quantified by ϵ_i. For each environmental configuration, a series of 30 trials have been repeated. The results described below report the scores of the trustors averaged on the whole series.

We refer to cognitive trustors for indicating the agents exploiting the capabilities to assess trust based on the cognitive architecture described in Section 4. Globally, six strategies are compared for the trustors:

Rand. Random agents adopt a random choice to decide the trustee to which delegate the task, based on a uniform distribution.

Stat. At each task completion, statistic agent stores the result value provided by the delegated trustee. Based on this information, at each round statistic agents compute trust for each trustee, by averaging the list of beliefs associated to each of them.

Cat. As described in Section 4, categorizing agents are able to prune the set of trustees assessing trust values based on the FCM mechanism. Cat evaluates the FCM on each trustee including its own beliefs related to abilities and dispositions with respect to the ongoing task. The leafs of the FCM are populated with the manifesta relieved on the trustee, and the weight of these connections is calculated using the function Ascribe. This FCM is also used as a basic cognitive configuration, which other cognitive agents are able to further refine.

Exp. Experience agents add to the FCM used by Cat a further branch resuming the past experience with the considered trustee for that task. The leafs of this branch are filled with the values coming form the belief base, which is updated in the same way as for Stat agent. Exploring set of possible trustees narrowed by the function Matches, the Exp agent is considerably more agile than the Stat agent in finding the best delegation option. On the other side, as for any other categorizer agent, if the optimal delegation involves an agent which is outside the set of categories filtered by the categorial reasoning, Exp agent will not be able to find it.

Ext. For any given task, the simulator provides a list of external conditions ($\mathcal{E}$) in which each trustee is assumed to carry out the possibly delegated task. External factors may refer to different environmental configurations, which may result in facilitating, being irrelevant or impeding conditions. Depending on the actual context, the reward of the task execution is improved, unchanged, or decreased. To consider external factors, Ext adds another branch to the same FCM used by Cat, which considers the context $\epsilon_i \in \mathcal{E}$ in which the trustee ag_i is going to execute

	Dental Operation		*Otitis*		*Appendicitis*	
Trustor	mean score	std	mean score	std	mean score	std
Rand	0.52	$\pm 10^{-1}$	0.5266	$\pm 10^{-1}$	0.48	$\pm 10^{-1}$
Cat	0.8565	$\pm 10^{-2}$	0.8479	$\pm 10^{-2}$	0.9050	$\pm 10^{-2}$
Exp	0.9782	$\pm 10^{-15}$	0.9753	$\pm 10^{-15}$	0.9857	$\pm 10^{-15}$
Ext	0.8569	$\pm 10^{-2}$	0.8511	$\pm 10^{-2}$	0.9043	$\pm 10^{-2}$
Stat	0.7615	$\pm 10^{-15}$	0.7625	$\pm 10^{-15}$	0.7442	$\pm 10^{-16}$
All	0.9782	$\pm 10^{-15}$	0.9753	$\pm 10^{-15}$	0.9858	$\pm 10^{-15}$

Table IV. Mean e standard deviation of the trustors' scores for the tasks *dental operation*, *appendicitis* and *otitis*, without environmental influence ($\delta = 0$).

the task. This context ϵ_i is stored as an internal belief in Ext's belief base, and is updated by perceiving the environment at each task assignment.

All. This cognitive agent adopts the complete set of information sources to assess trust. All utilizes a FCM including the manifesta of the trustee, that is its categorial information, as well as external factors (as Ext) and direct experience (as Exp).

In order to make the experimental results independent by the composition of the various population, the delegation effectiveness is measured in terms of absolute *score*. This metric is computed over the result of each delegation, and it is the ratio between the score actually obtained and the maximum available result achievable in the current population. In other words, a trustor obtains a score of 1.0 when it delegates the task to the trustee which is able to obtain the best performance among all the others.

The agents are implemented using Jason 1.3 [Bordini et al. 2007], while the simulator is based on CArtAgO 2.0 [Ricci et al. 2010], a platform for programming MAS environments based on artifacts. This choice allowed us to design agents in terms of epistemic and motivational attitudes, namely beliefs and goals. It also allowed the implementation of agents as driven by internal events, according to the programming model based on AgentSpeak(L). Such a programming model let us define the interaction between agents as based on messages, while the agent-environment interactions have been based on actions and perceptions, as defined in the JaCa programming model [Ricci et al. 2010][5]. The configuration of the machine on which the experiments have been run is: Intel(R) Core(TM) i5 CPU x64, 2.67 MHz, 6MB RAM, equipped with Windows 7, Java 1.6.

5.2 Results

Results are presented for the four cases of study where $\delta = 0, 5, 10, 20$ and the tasks are: *dental operation, appendicitis, otitis*.

5.2.1 $\delta = 0$ - *No influence.* Tab. IV reports the mean scores obtained by the trustors when the environmental influence is null (i.e., $\delta = 0$). The best performance is obtained by the cognitive trustors able to exploit experience of their past delegations: All and Exp (≈ 0.98 points for any tasks). Stat totally gets a lower ranking (≈ 0.75 points) although its score is lowered by the extensive learning phase during

[5]The complete distribution of the code for the employed agents, along with the results of the described experiments, are available at: `http://t3.istc.cnr.it/trustwiki/index.php/Cog-Trust`.

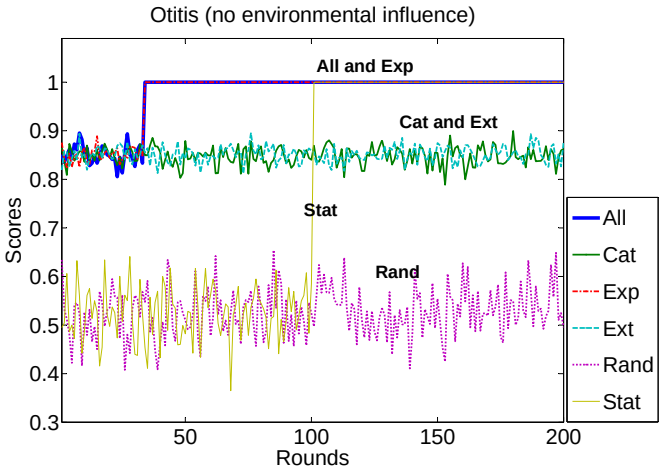
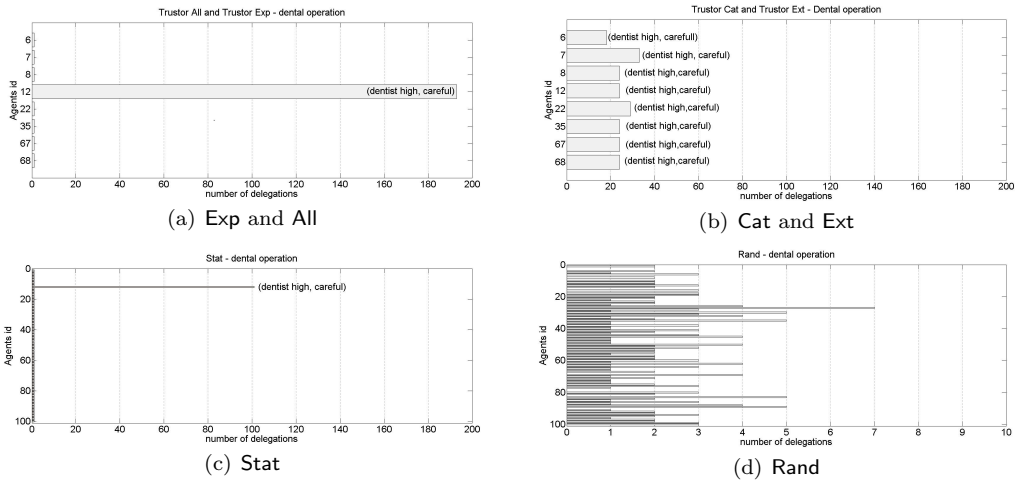


Fig. 2. Chart of the absolute scores of the trustors over rounds for the task *Otitis* without environmental influence ($\delta = 0$).

the first 100 iterations, needed to test at least one time each trustee. The task *otitis* is taken as instance in Fig. 2, showing the detail of the scores of the agents in each round. After a learning curve of only 25 iterations, Exp and All find the best performer in the trustees population and their scores stabilize on the maximum performance. Due to the cognitive attribution of trust based on categorization, the learning phase for these agents is limited to the exploration of only few trustees belonging to the most fitting categories for the task requirements. In a different way, the learning curve of the Stat agent requires a considerably larger number of iterations (100) before becoming stable on the maximum score. Cat picks up every trustee from the most fitting categories. Though, it is not able to further refine this choice not having any individual feedback from the delegated trustee. The score of this agent is ≈ 0.85 for *dental operation* and *otitis*, and ≈ 0.90 points for *appendicitis*. This difference is due to the different matching between categories and the task requirements: being *otitis* considered a very difficult task (see Tab. IV), the best available performance will be quite low and, in turns, the absolute score of categorizer agents will be proportionally higher. Ext performs the same as Cat because in absence of environment influence these two agents are in the same FCM configuration. Finally, Rand agents obtain the mean score (≈ 0.5 points) available on the whole set of performers.

Trust Delegation: It is also interesting to observe how the trustors distribute their delegations among the population of trustees: which trustee they delegate and how many times. The plots in Fig. 3 show the distribution of the delegations in a single experiment. In this case, ag_{12} is the best in performing the task *dental operation*, that is the trustee obtaining the best performance in the actual population and thus the reference agent on which the absolute score is computed. Stat delegates at least one time each agents,before finding and stabilizing on the best performer (Fig. 3(c)). Cat and Ext restrict their search to the only *good* and *available* dentists (Fig. 3(b)), but they are not able to identify the best performer among those. All and Exp,



(a) Exp and All (b) Cat and Ext

(c) Stat (d) Rand

Fig. 3. Experiment with no environment influence ($\delta = 0$). Delegations distribution

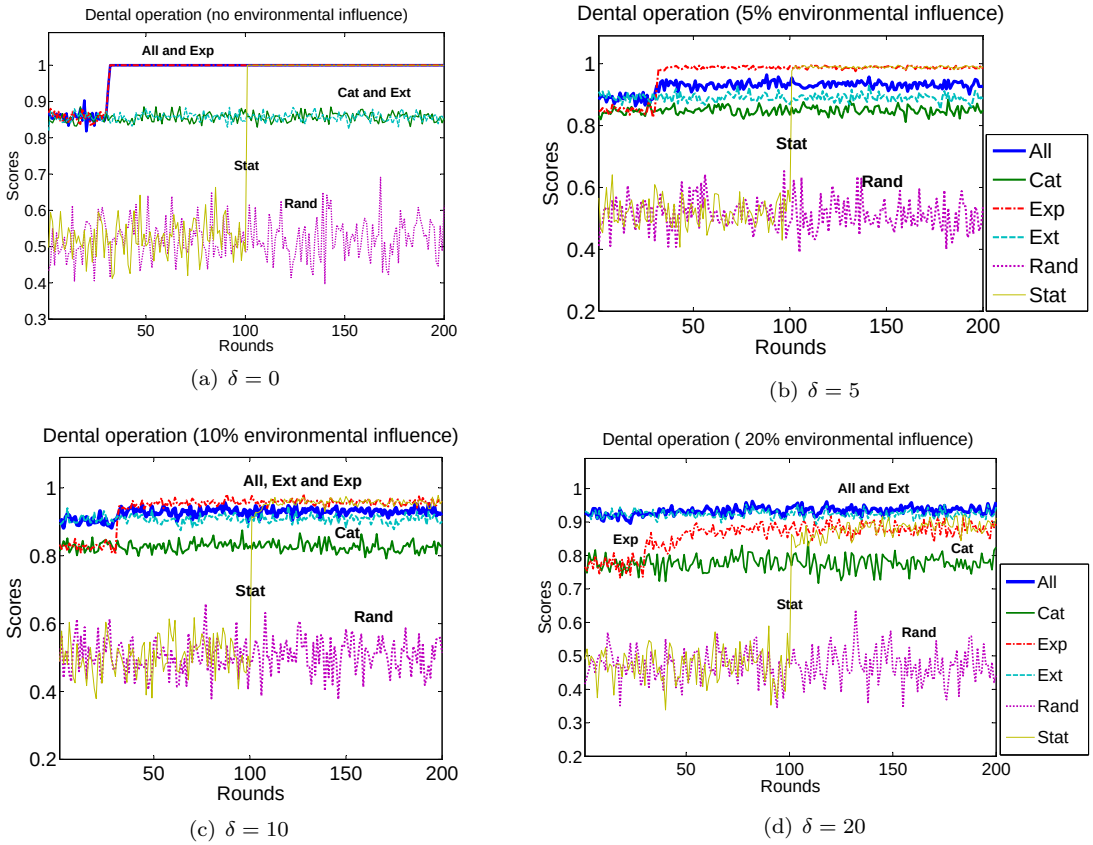
Dental operation

	$\delta = 5$		$\delta = 10$		$\delta = 20$	
Trustor	mean score	std	mean score	std	mean score	std
Rand	0.5125	$\pm 10^{-1}$	0.5047	$\pm 10^{-1}$	0.4711	$\pm 10^{-1}$
Cat	0.8489	$\pm 10^{-2}$	0.8272	$\pm 10^{-2}$	0.7769	$\pm 10^{-2}$
Exp	0.9669	$\pm 10^{-2}$	0.9368	$\pm 10^{-1}$	0.8592	$\pm 10^{-2}$
Ext	0.8904	$\pm 10^{-2}$	0.9080	$\pm 10^{-2}$	0.9233	$\pm 10^{-1}$
Stat	0.7533	$\pm 10^{-2}$	0.7294	$\pm 10^{-2}$	0.67555	$\pm 10^{-2}$
All	0.9256	$\pm 10^{-2}$	0.9263	$\pm 10^{-2}$	0.9328	$\pm 10^{-2}$

Table V. Mean e standard deviation of the trustors' scores for the task *dental operation*, varying the environmental influence: $\delta = 5, 10, 20$.

combining together categorization and direct experience, quickly identify the best trustee in the group and delegate it throughout the simulation (Fig. 3(a)). Finally, Rand delegates uniformly amongst the whole trustees population (Fig. 3(d)).

5.2.2 $\delta = 5$ *and* $\delta = 10$ - *Low and Medium influence.* For the sake of simplicity we focus here on the only task *dental operation*, although similar analysis can be extended to the other tasks. Tab. V reports the scores for the same set of trustors with low ($\delta = 5$) and medium ($\delta = 10$) environmental influence. Strategies considering a richer set of information sources are expected to outperform more simple strategies. Compared to the case of $\delta = 0$, the standard deviation of the average results is sensibly increased (from 10^{-15} to 10^{-2}) as a sign that the environmental noise sheds more uncertainty over the system. Increasing δ, Ext raises its score up to 0.90 while other cognitive agents Exp and Cat loose around 3% and 4%. For $\delta = 10$, the environmental influence is strong enough to counter the direct experience and Exp and Ext equalize their scores on ≈ 0.91. The noise provoked by the environment on the performance is remarked by some irregularities on the curve of Exp, showed in Fig. 4(b) and Fig. 4(c). In this configuration, the trustor who delegates always to the same trustee will receive different outcomes from round to



(a) $\delta = 0$ (b) $\delta = 5$

(c) $\delta = 10$ (d) $\delta = 20$

Fig. 4. Scores of the trustors for the task *Dental operation*, varying the environmental influence $\delta = 0, 5, 10, 20$.

round, due to the environmental noise. Using the entire set of information sources, All builds its delegation choice on both experience and environmental factors, Curiously, when $\delta = 5$, All does worse than Exp (0.92 vs. 0.96). To explain that, we have to remark that we could have e-factors overinfluencing the degree of trust and thus do not leading to the correct anticipation of the trustee's performance. Different dynamics can be obtained refining the tradeoff between the influences of experience and e-factors. Anyways, All is able to maintain its score around 0.92 in both the environment configurations. Stat and Cat agents show a general decrease of their scores when δ increases. Cat scores 0.84 with low influence and 0.82 with medium influence. Stat steadies on the best trustee after having completed a full learning phase and gets the scores: 0.75 (low influence) and 0.72 (medium influence). Finally, Rand attains the worst performance, with a mean score around 0.50 points.

5.2.3 $\delta = 20$ - *High influence.* When the influence on task execution is high ($\delta = 20$), the noise of the environment undermines every direct experience (see Fig. 4(d)). This fact fosters the capability of the agents to exploit the information

Cancer	
Abilities	
manuality	90
radiotheraphy_spec	99
oncology_spec	99
expertise	90
problem solving	90
Dispositions	
availability	80
caution	90
attention	90

Infection	
Abilities	
manuality	70
expertise	70
problem solving	70
Dispositions	
availability	60
caution	70
attention	70

Table VI. Tasks with double (left) and missing (right) specialization requirement

	Cancer		*Infection*	
Trustor	mean score	std	mean score	std
Rand	0.5836	$\pm 10^{-1}$	0.6840	$\pm 10^{-1}$
Cat	0.8248	$\pm 10^{-2}$	0.9212	$\pm 10^{-2}$
Exp	0.9622	$\pm 10^{-16}$	0.9198	$\pm 10^{-15}$
Ext	0.8242	$\pm 10^{-2}$	0.9212	$\pm 10^{-2}$
Stat	0.7888	$\pm 10^{-16}$	0.8413	$\pm 10^{-15}$
All	0.9622	$\pm 10^{-16}$	0.9198	$\pm 10^{-15}$

Table VII. Mean e standard deviation of the trustors' scores for the tasks *cancer* and *infection*, without environmental influence ($\delta = 0$).

of the external factors. Results show that All is again resistent to the environmental influences and keeps its score on 0.93. Ext, only following up the trustee with the most positive context (i.e., considering a much smaller trust sources set), performs nearly the same as All (0.92) and is now able to do much better than Exp (0.92 vs 0.85). Again Stat and Cat keep loosing points because they are not able to anticipate environment influences. No differences are observed for Rand agent, whose performance in fact does not depend on the environment conditions.

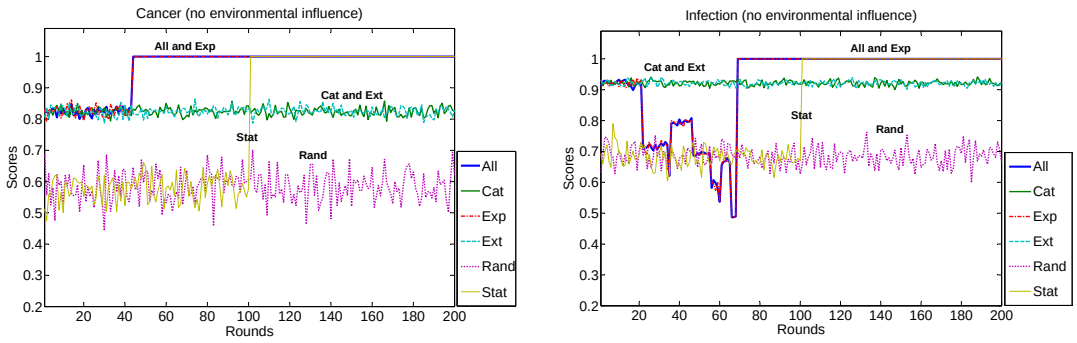
5.3 The role of professionalizing features

So far we dealt with tasks featuring a specific professional category, i.e. tasks containing one and only one discriminant requirement (e.g., *dentistry_spec* for *dental operation*). This assumption is justified by common knowledge or personal experience, suggesting that tasks are often a priori assumed to be concerned with some professional specialization. For instance, in our scenario the agents know that a dental operation can only be fulfilled by dentists, as well as appendicitis by surgeons and so forth. In the categorial reasoning, this discriminant requirement has a pivotal role, as it allows for immediately cutting off the professional categories which are considered to be *unsuitable* for a given task. We are now interested in deeper investigating the role of the specialization requirement in our computational model. In particular, we address the cases in which the specializing requirement might be over specified or missing. For this purpose we consider two new tasks, *Cancer* and *Infection*: the former with *double* discriminant requirements, the latter *without* any discriminant requirements (Tab. VI).

Cancer: Tab. VII shows the results of the six trustors playing on the task *Cancer*.

The double specialization requirement has the effect to *extend* the set of suitable professional categories to the *oncologists* and the *radiotherapists*. This increases the probability to find a good (but not the best) performer for strategies with random search such as Rand and Stat (the latter only in his learning phase). As a consequence, all the trustors improve their scores and the gap between categorizer agents and the other strategies is reduced. We can notice in Fig. 5(a) that Ext and All are involved in a larger exploration of approximately 45 iterations. The structure of this task is also useful to understand the role of the function Matches (Algorithm 3) in the computational architecture of our model. Since we assume there are no categories having both the two specialization requirements specified by *Cancer*, the matching values of any professional categories on this task computed by Ascribe will be lower. In other words, the analysis features-requirements so far performed is not able in this case to clearly point out the most suitable professional categories for this task to the eyes of the agent. As a consequence, the agents will need to use a lower value of pruning threshold in the function Matches in order to consider a sufficient number of professional categories in their search space.
Infection: For the task *infection no* specialization requirement is specified. Ideally, this task represents a sort of abstract task for which the agent *does not know* exactly the concerned doctor specialization to cope with it. The lack of the direct relation between task and a professional specialization opens up a number of alternatives in the agent's choices: we might be in the opposite situations in which either many professional categories could be able to successfully carry out *Infection* or nobody actually can perform well in the current population. Results in Tab. VII point out a global flattening in the performances of all the trustors scoring around 0.9 with no significant differences. This is due to the absence of a trustee who is able to outperform the current population on such a task, so statistical and the categorial strategies cannot do any better than delegate some random trustees. Fig. 5(b) also shows that Exp and All are forced to get some poor delegations up to 0.5 points as a consequence of their blind exploration. When the agent is not able to immediately refer the task to some specializations, the categorial reasoning does not add any substantial help to the search of the best trustee although still categorizer agents tries (and in fact they do) to attribute a most suitable category to the given task. In the end, Cat and Ext are still able to maintain a slight advantage against Rand and Stat, with 0.90 against 0.84.

These examples allow to generalize the applicability of our approach. Overspecialized tasks like Cancer, which present requirements drawn on several specializations, force the agent to explore a wide range of professional categories. Besides, if task is non-specialized, and if the task specification does not reflect the taxonomy of the categories like in Infection, then the categorization is not able to exercise any pruning on the population of trustees and agents need to try many delegation options. A relevant result can be analyzed in the trend of agents exploiting experience All and Exp. Having no environment influences ($\delta = 0$), the impact of external factors is inhibited in All, and it exploits the same TrustEval function adopted by the Exp agent. In the Cancer task, the learning phase is doubled wit respect to the one done in the previous tasks. The exploration now reaches about 50 rounds, and can be explained by the fact that now All and Exp agents include the exploration



(a) Cancer (b) Infection

Fig. 5. Chart of the scores of the trustors over rounds for the task *Cancer* (Left) and *Infection* (Right) with no environmental influence ($\delta = 0$).

of both *Radiotherapists* and *Oncologists*. In the Infection task, the learning phase is tripled, it reaches 75 rounds and clearly includes the exploration of all the encountered categories. This can be explained since the fact that the Infection task does not specify a diriment requirement to be matched with the professionalizing features of the categories. The exploration phase of All and Exp proceeds category by category. As shown in Fig. 5(b), their scores undergo a stepped trend, each step representing the mean score provided by one single explored category.

5.4 Discussion

The experimentation has been extensive and has considered an heterogeneous sample of tasks, categories and environmental conditions: more than 600 simulations were performed. The distinctive feature of the cognitive trustor is the capability to develop a reasoning over multiple levels: *personal, categorial, contextual*. Personal level takes into account the *direct experience* of a particular trustee. Categorial level considers *abilities* and *dispositions* of the potential trustee on the basis of the categories it belongs to. Finally the contextual level takes into account the specific *environmental conditions* in which the trustee is going to realize the task. The FCM structure is an image of such a hierarchical organization of the information sources.

Results pointed out an overall superior performance of the categorial agents against the mechanism of pure statistical trust. Statistical agent is forced to explore the whole population, thus requiring a huge (and computationally hard) amount of resources to find the best performer. Categorizing agents are able to ease the decision choices by pruning trustees belonging to the unsuitable categories. In doing this, the cognitive model bridges a gap of knowledge: it allows the trustor to infer the effective abilities of a trustee (krypta), namely forming an expectation about its possible performance, based on categorial analysis of observable knowledge (manifesta). The combination of categorial reasoning and direct experience allow for a search bounded to the only appropriate trustees and a consequent drastic reduction of the learning time. After the learning stage, both the strategies stabilize on the same performance.

Environmental conditions represent a further aspect to refine and enhance the cognitive trust reasoning. By considering not only agent's skills but also the influence of the environment on the task fulfillment, the cognitive agent Ext foresees situations of a poor outcome from apparently good performer and instead foster delegation to trustees not only skilled but also working in a favorable environment. By varying the δ parameter we observed the advantage of agent Ext compared with the other strategies (case $\delta = 20$). Nevertheless, in some cases an expected positive environmental influence could not be enough to find the optimal delegation, when the performance gap between agents holding the same category is high (case $\delta = 10$).

Finally, we discussed how the task specification directly affects the search domain of the categorizer trustor, with a special focus on the professionalizing features. Professionalizing features help the categorization process as they allow for an immediate cut of a part of unsuitable categories. Anyways, categorization relies on a wider spectrum of features (see for example *dispositions* for which specializing features are not considered) and the performance of categorizer agents only partially depends on whether or not the specializing feature occurs. It has to be remarked that both the choices of task requirements and categorial intervals is made off-line, and it is due to designer preferences. For instance, one could define *caution* for the *Careful* category in the interval [75, 95] instead of [80, 100]. As experiments show, this would compromise only partially the ability of cognitive agents to assess trustworthiness and find the best trustee. Categories are indeed evaluated with respect to the task. The role of the Ascribe function is crucial in centering the categories with the requirements to fulfill. Although the reasoning provides a slightly different set of possible trustees to choose in, it still allows agents to maintain similar results in terms of delegation effectiveness. Is worth nothing that the natural progress of a MAS, where agents are assumed to learn and continually enrich their knowledge, brings about the case in which the relation between tasks and specialized categories is known and can be suitably exploited for evaluating trust. This suggests an important scalability for the categorial approach: the model can be effectively applied in a wide class of domains, where it is feasible to associate specializing requirements to classes or groups of agents.

6. CONCLUSIONS AND RELATED WORKS

In this work we started from the fundamental idea that there is a strong relationship between the uncertainty in trusting agents and the fact that part of the qualities of a possible trustee is unobservable. Based on this assumption, we provided a computational model by which cognitive agents succeed to assess trust upon a population of heterogeneous and possibly unknown trustees. As showed, the proposed approach enables a particular kind of reasoning, which is based either on direct experience of past interactions (personal dimension) either on information about professional and dispositional abilities which can be assumed on generic and open population of agents. Conducted experiments clearly show the benefits of managing this twofold heuristic, thus effectively improving delegation strategies under uncertainty and ameliorating tasks fulfillment with respect to traditional strategies based on only direct experience.

As in Bacharach and Gambetta [Bacharach and Gambetta 2001], we realized this ability of dividing the information and characterizing the system in "krypta and manifesta", thus enabling a special kind of inference allowing to explain agents' internal qualities (krypta) with their observable signs (manifesta). In particular, we extended this relevant concept to the categories: an agent expressing the signs of a category, inherits the qualifying properties of that category. Our results show that categories result as a pivotal piece of information for agents who are able to manage it. Indeed categorial reasoning allows to establish fruitful interactions with agents which have not been encountered yet. This aspect has a important significance in the context of open-systems, characterized by heterogeneous societies of self-interested agents. Our model makes the MAS open to new agents that can enter and leave the application at any time, but not open to categories, that are in fact preexisting for the agents. Our future work will be addressed at addressing this limitation, by developing mechanisms to fully enable the categorial reasoning in open systems, for instance letting categories to emerge on the basis of individual experience. A similar approach has been recently developed by Burnett, Norman and Sycara [Burnett et al. 2010], where the categories or class are not preexisting to the interactions, but the agents can generalize their experiences with known partners in previous contexts. This work shows that, by using data-mining techniques, agents can form stereotypes that allow to bootstrap trust evaluations about unknown agents in new contexts. By ascribing trust evaluations to learned classes of individuals as well as individuals themselves, agents can make use of both previous experiences and reputational opinions in contexts where this would not otherwise be possible.

Another limitation of our model is the naive statistical model that we introduced in order to benchmark and evaluate the approach. Among other weakness, the statistical approach that we implemented is not able to recognize environment influences. More in general, there are several interesting studies analyzing the role of the context in the trust relationships [Rehak et al. 2007], [Tavakolifard et al. 2008]. In these works the constraints introduced by the context and its various dimensions are formally and/or informally analyzed as a support for evaluating trust relationships. These studies can be considered as partially overlapping our approach, even if in our model we focus explicitly on categories (just eventually one of the constraints in those works) as a fundamental instrument for attributing features and properties to agents.

Finally, we have already implemented a fuzzy approach [Falcone et al. 2003; Falcone et al. 2004] of our socio-cognitive model of trust. This paper revised and extended the conceptual analysis of that work in modeling the fuzzy cognitive maps used by the cognitive agents. The next steps in this direction will be in the development of a modular architecture, i.e., able to configure the topology of the FCMs based on the relevant aspects characterizing particular situations. Future works will also account the possibility for agents to improve and refine the categories at runtime, thus based on the situated conditions of the system. A further step will account the possibility for agents to share such refined information, thus enabling different kinds of cooperation based on the communication of the categorial information characterizing the system. In any case the aim of the present work was

to verify the relevant role of the category-based analysis for trusting agents. On this basis we have adapted and updated the cognitive maps also introducing the role played by the categorial structures and by the experience.

REFERENCES

Bacharach, M. and Gambetta, D. 2001. Trust as type detection. In *Trust and deception in virtual societies*. Kluwer Academic Publishers, 1–26.

Barber, K. and Kim, J. 2001. Belief revision process based on trust: Agents evaluating reputation of information sources. *Trust in Cyber-societies, Lecture Notes in Computer Science (LNCS) 2246/2001*, 73–82.

Bordini, R. H., Hübner, J. F., and Wooldrige, M. 2007. *Programming Multi-Agent Systems in AgentSpeak using Jason*. Wiley Series in Agent Technology. John Wiley & Sons.

Bueno, S. and Salmeron, J. 2009. Benchmarking main activation functions in Fuzzy Cognitive Maps. *Expert Systems with Applications: An International Journal 36,* 3, 5221–5229.

Burnett, C., Norman, T., and Sycara, K. 2010. Bootstrapping trust evaluations through stereotypes. In *9th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, van der Hoek, Kaminka, Lesperance, Luck, and Sen, Eds. Toronto, Canada, 241–248.

Castelfranchi, C. and Falcone, R. 1997. Delegation Conflicts. In *Multi-Agent Rationality*, M. Boman and W. Van de Velde, Eds. Lecture Notes in Artificial Intelligence, vol. 1237. Springer-Verlag, 234–254.

Castelfranchi, C. and Falcone, R. 1998. Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification. In *Proceedings of the 3rd International Conference on Multi Agent Systems*. 72–79.

Castelfranchi, C. and Falcone, R. 2010. *Trust Theory. A Socio-Cognitive and Computational Model*. Wiley Series in Agent Technology. John Wiley & Sons.

Falcone, R. and Castelfranchi, C. 2002. Social Trust: a Cognitive Approach. *Trust and deception in virtual societies*, 55–90.

Falcone, R., Pezzulo, G., and Castelfranchi, C. 2003. A fuzzy approach to a belief-based trust computation. *Trust, reputation, and security: theories and practice*, 55–60.

Falcone, R., Pezzulo, G., Castelfranchi, C., and Calvi, G. 2004. Why a cognitive trustier perform better: Simulating trust-based Contract Nets. In *3rd Int. Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS-04)*. ACM, 1392–1393.

Hang, C., Wang, Y., and Singh, M. 2009. Operators for propagating trust and their evaluation in social networks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. IFAAMAS, 1025–1032.

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

Jonker, C. and Treur, J. 1999. Formal analysis of models for the dynamics of trust based on experiences. *Multi-Agent System Engineering*, 221–231.

Kosko, B. 1986. Fuzzy Cognitive Maps. *International Journal of Man-Machine Studies 24,* 1, 65–75.

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

Marsh, S. 1994. Formalising trust as a computational concept. Ph.D. thesis, University of Stirling. PhD thesis.

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

Rehak, M., Gregor, M., Pechoucek, M., and Bradshaw, J. 2007. Representing context for multiagent trust modeling. In *Intelligent Agent Technology, 2006. IAT'06. IEEE/WIC/ACM International Conference on*. IEEE, 737–746.

Resnick, P. and Zeckhauser, R. 2002. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. *Advances in Applied Microeconomics: A Research Annual 11*, 127–157.

Ricci, A., Piunti, M., and Viroli, M. 2010. Environment Programming in Multi-Agent Systems: An Artifact-Based Perspective. *Autonomous Agents and Multi-Agent Systems*. Published Online with ISSN 1573-7454 (will appear with ISSN 1387-2532).

Sabater, J. 2003. Trust and reputation for agent societies. Ph.D. thesis, Universitat Autonoma de Barcelona.

Tavakolifard, M., Knapskog, S., and Herrmann, P. 2008. Trust transferability among similar contexts. In *Proceedings of the 4th ACM symposium on QoS and security for wireless and mobile networks*. ACM, 91–97.

Yolum, P. and Singh, M. P. 2003. Emergent Properties of ReferralSystems. In *Proceedings of the 2nd International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*.

Yu, B. and Singh, M. 2003. Searching social networks. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. ACM, 72.

Ziegler, C. 2009. On Propagating Interpersonal Trust in Social Networks. *Computing with Social Trust*, 133–168.