MODELLING THE MIND: STUDYING THE DECISIONAL PROCESSES BY THE MEANS OF TRUST ON INFORMATION SOURCES

Author: Alessandro Sapienza

Advisor: Prof. Corrado Santoro

1. INDEX

ACKNOWLEDGMENTS

Many people supported me during my thesis work. I would like to thanks whose contribute was indispensable.
First of all, I am grateful to the Prof. Corrado Santoro, the Dr. Rino Falcone and the Prof. Cristiano Castelfanchi, which helped me growing in my job, always giving me new inputs and hints.

Nevertheless, I intend to thank my family for supporting me in these years outside of my job, believing in me, trusting and supporting me even in the difficult moments.

INTRODUCTION

In every moment, the events that happen in our life put us in the condition to make a decision. Maybe this decision could refer to a not so much relevant task ("today I need to go to work, should I take the bus or the car?", "should I see the film with Mario or should I have a dinner with Carla this Saturday?"), so that it is relatively easy to decide, probably because each possible choice has not a great impact on our life: we are not going to take a risk (or there is a low level of risk), there is not a heavy possible lose that threatens us.

But there can be situations in which the choice we take can seriously have a great impact, and then require a bigger cognitive effort. Sometimes we have to take into account a risk factor; there could be the possibility to lose material goods or even to compromise our health itself. What to do in these risky situations? As it is easy to understand, in these situations the decisional process becomes heavier; identifying a correct choice or at least the best one is a critical task.

In order to make a decision, it is necessary a proper quantity of knowledge. The first thing to do is quite obvious: we try to use the knowledge we already have, using the one that represents our basic beliefs, the knowledge that we have about the world in which we live. It has been built and consolidated in time and it is available each time we need it.

This knowledge we possess and we can immediately access actually has been previously produced in various ways. Sometimes it derives from our direct experience and we are able to acquire and evaluate it though our senses/reasoning. Sometimes we acquire it from external sources and we can (or not) remember how we acquired it and how much these sources are reliable. In any case we consider this knowledge as our own. This ability to reason on knowledge built in time represents an extraordinary modality to evaluate the world.

However, in many case our own knowledge is not enough for the task that we have to face in the real world. For instance we could have the necessity to make a decision in a context in which we are not competent, or maybe we do not have the possibility to evaluate properly the situation. Maybe we could find our self in a dynamic system, in which it is not so easy to make a forecast of how it will change.

In conclusion, sometimes to use just our knowledge is not enough. And that is fine, natural. It is a necessary came out with the same modern society. The knowledge has become so large, wide that no human being is able to handle it completely, but it can manage only a portion of it. Because of that, it is possible to become a master in some subject, and even then to be a complete ignorant in some others. This is not a negative effect, on the contrary; it is a necessity, a natural evolution. How Thomas Sowell states (Sowell, 1980):

“individually we know so pathetically little, and yet socially we use a range and complexity of knowledge that would confound a computer”.

Then it seems that human beings find in the social source the knowledge they need to manage their problems. Therefore in many situations that occur in our daily life it is necessary to ask someone/something else, to use knowledge that comes from external sources. It can be a human source, but this is not necessarily true (a book, a web page, our personal smartphone and so on). In any case, having more information allows us to decide better.

However when we try to use information coming from an external source, we get into some complicated dynamic: it rises up some problems that it is necessary to face.

First of all, how can I be sure that the information I use is true? It has been reported to me by a given external source, but should I trust this source? Who does ensure me that this source is able to evaluate the specific situation better than me? Moreover, should I believe that this source is not willing to deceive me, that it is not intentionally trying to make me choose wrongly? Maybe the source sent me correctly its information, but my I trust the communication channel? Has the information been altered by external influences

Seeing this problem from another perspective, I need to be sure of my abilities. Maybe I am the one who did not understand correctly the message or simply does not recall properly its content.

In this study we intend to understand which are the methodologies that allow using effectively external information sources for our internal decisional processes, with the aim of identifying a choice between a series of alternatives. In particular, we propose the use of the concept of trust applied to the information sources.

In order to do so, we first developed a theory about trust on information sources, starting from the classical model of trust already proposed in literature. The one of trust is the key concept that allows us to use information sources. Thanks to it, we are able to consider differently each source, understanding which are more trustworthy and then can be taken into account while reasoning and those which are not enough trustworthy that has to be treated differently (at least, they must have a lighter weight). Then trust becomes a weight to use towards information sources.

Even if trust can derive from a reasoning process, it can also be irrational or impulsive. I can trust (or distrust) someone just because my feelings tell me to do it. I can feel affection for someone (i.e. a relative) and this affection can influence my rational judgment.

The same stands for moods. Mood is a property that depends just on the trustor; it does not concern the relationship with the trustee. But similarly to emotions, it can affect the final trust decision. A lot of studies in literature examine this topic (Capra, 2004)(Myers and Tingley, 2011)(Sharpanskykh and Treur, 2010).

However in this work we are not going to discuss this form of trust, as we will focus on its rational part.

After a pure theoretical study, we implemented our theory at a computational level, producing a series of models, each of which is the evolution of the previous one as it allows more expressive power.

Then, by the means of simulative technologies we studied the practical aspect of this theory, applying these computational models to practical context. This allowed us to identify some interesting aspect both from the theoretical level and the practical level.

The rest of the work will be articulated as follows:

1. Chapter 1 describes the state of the art about trust and information sources

2. In Chapter 2 we introduce the computational models we used in our work.
3. In Chapter 3 we investigate the role of categorizes for trusting information sources.
4. Chapter 4 proposes an analysis about recommendation; in particular we test the concept of category's recommendation, comparing it to the classical concept of individual recommendation.
5. In Chapter 5 we analyze the role of information sources in case of critical scenarios, such us those of critical hydrogeological phenomena.
6. Chapter 6 summarizes the whole work and its results.

CHAPTER 1:

STATE OF THE ART

This chapter explains the current situation about trust and in particular trust on information source. A lot could be said about that, especially on trust. But in order to give more focus on the main work, I decide not to expand too much this section. Furthermore, a lot of discussion about the state of the art can be found later in each specific work.

1. The concept of Trust

Talking about trust would require a lot of time and space; it is not an easy task, since it contains a lot of aspects of social (and non social) interaction. However, in order to pursuit our work, we just need to understand how this context is used in literature and then to identify which could be the best option according to us, so that we can effectively exploit it in the context of information sources.

As (Mutti, 1987) argues "the number of meanings attributed to the idea of trust in social analysis is disconcerting. Certainly this deplorable state of things is the product of a general theoretical negligence. It is almost as if, due to some strange self-reflecting mechanism, social science has ended up losing its own trust in the possibility of considering trust in a significant way".

Trust is in fact a complex concept to define. The primary reason of this difficulty is that trust is a generic concept that can be applied to multiple fields: psychology, sociology, computer science, economics, and so on. So it is natural that the members of each community tried to provide a definition of trust from their specific point of view.

As an example of limitative definition, Barber and Kim (Barber and Kim, 2001) define trust as the agent's confidence in the ability and intention of an information source to deliver correct information. This is a very good definition, but it is so specific that its validity is strictly related to the context of information sources (that is the one of their work).

A base truth for each definition is that a trust relationship is between two elements: a trustor and a trustee; the trustor trusts the trustee. We are going to see some interesting trust definition, analyzing how they describe the relationship between trustor and trustee, in order to understand why they work or not.

Let's start from the definition provided by the online Cambridge dictionary (http://dictionary.cambridge.org/dictionary/english/trust): to trust means "**to believe that someone is good and honest and will not harm you, or that something is safe and reliable**". Actually, this is a very limitative definition. It starts from the point of view that trust is a relation between two people, completely excluding the possibility to trust something. However, trust requires that the trustor is a cognitive agent, but there is no constrain on the trustee's nature. For instance, I can trust the floor when I walk as I hope that it will not break. Moreover I can trust my car when I turn it on for carrying me where I want. In other words, trust towards objects is still a kind of trust.

But this is not the unique problem with this definition. Believing that someone will not harm me is a good starting point to trust him, but it is not trust: a belief is just a belief. Trust involves also the "act" of trusting.

A second definition is the one proposed by Gambetta (Gambetta, 2000). According to him "trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action [...]. When we say we trust someone or that someone is trustworthy, we implicitly mean that the probability that he will perform an action that is beneficial or at least not detrimental to us is high enough for us to consider engaging in some form of cooperation with him."

Even if Gambetta reduces trust to a mere probability (losing a lot of interesting aspect), this is indeed a good trust definition, strongly accepted in literature. However it focuses just on one part of trust: the belief, the evaluation.

Similarly to Gambetta, (Sztompka, 1999) focuses on the probabilistic nature of trust. According to him "trust is a bet about the future contingent actions of others". In particular with this definition, the author wants to point out two main aspect of trust: beliefs and commitment. Trust is in fact based on the specific belief that someone will do something in the future. It is like if using trust we are anticipating the future, making a forecast.

He also identifies that this anticipatory belief is not enough to trust someone. It is also necessary to commit ourselves to some action that has an unsure outcome. Then in order to trust the commitment is necessary.

This is very interesting as to bet and to be committed have one thing in common: the risk. As I bet on someone actions, as I am committed, I am taking a lot of risks:

1. The first risk is that the goal is not achieved, maybe because the trustee is not competent or it is malevolent.
2. Given that I trust the trustee, I am exposed to it. I will not be suspicious and I will not monitor him.
3. Then I risk to lose what I invested. For instance, I could have paid the trustee in order to pursuit the task.
4. I could lose the possibility to achieve my goal: maybe I will not be able to find someone else to pursuit this task or the task itself is not repeatable.

In (Mayer et al, 1995) authors believe that trust is "the willingness of a party (the trustor) to be vulnerable to the actions of another party (the trustee) based on the expectation that the trustee will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party".
They focus on the decisional phase, in which an agent decides to trust someone/something, pointing out that this decision makes him (consciously) vulnerable to another agent.

Castelfranchi and Falcone (Castelfranchi and Falcone, 2010) provide in their cognitive model of trust a very complete and detailed definition. According to them, trust is a concept that has to be developed into several levels:

- As an attitude or disposition (both cognitive and affective) towards the trustee;

- As a decision and intention based on the previous disposition;
- In the end, as the action of trusting that implies a relationship with the trustee.

Let's consider a context in which a cognitive agent X is interested in pursuing a specific goal and it has the necessity, for that specific goal, to delegate a task to another agent Y (cognitive or not). In order to carry out the delegation, it is necessary a certain level of trust towards Y, and this trust is based on a preliminary evaluation of its internal feature.

The one proposed by Castelfranchi and Falcone seems to be the most complete definition of trust, as it is context-independent and it allows an analysis of all the phases of trusting (evaluation, decision and act). Plus, they clearly identify that trust is not a vague concept, but it is always linked to a task: I don't trust someone/something; I trust him/it for a task (which can be specified, abstract of even a category of task). An agent can be trustworthy for a task but completely untrustworthy for another one. The reason of its trustworthiness has to be searched in features of competence and willingness of the agent (which again are related to the specific task).

Being the most complete and exhaustive, we decided to use this last model as a base for our work.

2. Trust on Information Sources

In the previous paragraph we focused on the general concept of trust. We started analyzing a number of definitions proposed in literature, seeing their strength and weakness. In the end, we decided to use the cognitive model of trust Castelfranchi-Falcone, as we believe that it is the most useful in our case.

However we are interested in the specific concept of trust on information sources. This is not a novel approach in literature, as many other researchers tried to provide a solution to the problem of information source using trust. We are going to propose a specific analysis of the works in literature addressing this problem, so that we can understand how they exploit trust to estimate the uncertainty on the information.

Trust in information sources: seeking information from people, documents, and virtual agents

In (Hertzum et al, 2002) the authors propose an interesting analysis of trust on information sources. In their work they investigate why trust is important for selecting information sources, as the trustworthiness of a source is strictly related to the perceived quality of the information itself. In particular, they show that "software engineers and users of e-commerce websites devote a lot of attention to considerations about the trustworthiness of their sources, which include people, documents, and virtual agents."

Concerning trust, they state that in order to use an information source, I don't just need to identify it as trustworthy but I also must know that I am able to understand what the source is reporting. It means that people look for information that it is not just accessible, but presented in a comprehensible way, so that it is possible to evaluate its quality. This is in fact a necessary step, as the quality of information is not just given, but it is subjectively evaluated.

They state that "previous work on the factors that enter into people's assessment of people, documents, and IT systems as sources of information has been conducted under the headings of the:"

1. least-effort principle: it is a value coming from the combination of cost (in term of time effort or resource) and quality of the information (King et al, 1994) (Pinelli et al, 1993);
2. relevance: usually people don't look for information in general, but relevant to specific task, context or problem;
3. usability: this is actually an ambiguous term that has different meanings.

Starting from this assumption, the authors made two experiments to study the role of the perceived trustworthiness of the information source. The first one concerns how engineers evaluate and then select people as information sources.

The research was conducted analyzing the discussion of a team of engineers working on a project. The results of this experiment show that, while selecting information sources, they paid much more attention to the quality of the source rather than to the cost of using it. This result was particularly evident when the sources were people, but it is still evident for documents.

In particular, in this first experiment it emerges that engineers preferred as information sources people that they knew directly or that were recommended by a colleague.

In the second experiment, the authors investigate how users interface with virtual personified agents. A personified agent is an anthropomorphised agent, which has a human aspect. In the specific context, the users navigate in a virtual store and the aim of the virtual agent is to help users in their navigation so that they can find easily what they need. It is also proactive, meaning that it will try to provide suggestions in advance.

While users were interacting with the virtual agent trying to solve some given tasks, a researcher was interviewing them to understand their behavior and their choices. After that, it followed a focus group session. Trying to identify the requirements that a personified virtual agents should have, it was clear that trustworthiness was the most important and crucial one. Plus, many of the other requirements were close to the concept of trust.

In conclusion, it seems quite clear that, both interacting with human or virtual agents, trust is a key factor to accept them as information sources.

Trust on Beliefs: Source, Time and Expertise

In (Melo et al, 2016) authors extend the concept of trust to agents' beliefs, combining the provenance of the information (the source) with how much the information is up to date. In particular, in this specific approach authors provide the possibility to give more or less weight to one of these two meta-information: the source's evaluation or to the information freshness. According to that, they propose different agent's profile, which use differently these two dimensions.

Authors also show how patterns of reasoning such as argumentation schemes can play an important role in their approach, considering the expertise of the source of information.

It is worth noting how they define and use trust. According to them trust is a relation between two agents and it is asymmetric; the fact that X has a trust relationship with Y with degree t does not mean that the same hold for Y: it could trust more or less X or it could not trust it at all. What is seems to lack is an explicit link with a task: trust is always relative to a task; I am not going to trust someone for everything, but just for something.

Authors also point out the difference between having a trust degree equal to 0 and not having a trust degree. The first one means that the trustor completely distrusts the source; the second one means that the trustor has no evaluation for the source, it is not able to evaluate it.
This is clearly a valid concept, but I personally think that using a null value is not the better solution, as it creates confusion in the computational process (how do you compute or consider a null value?). A better and much more elegant solution would have been to use a value of 0.5, which actually is a neutral value (between the complete distrust of 0 and the compete trust of 1) and indicates a situation of uncertainty.

It is very interesting the concept of trust network, represented by the means of a direct graph. Here each agent is represented with a node and the directional edges that connect two agents represent trust relationship. An edge starting from agent Ag_i and ending in agent Ag_j means that Ag_i trusts Ag_j with a degree of trust given by the label of the edge.
In this graph there are two kind of edge. Those representing direct trust and those representing indirect trust.

However, as a consequence of not considering the task in the trust evaluations, authors make a logical mistake in introducing indirect trust.
Suppose that Ag_i did not have any experience with Ag_j, it cannot directly evaluate it. However they state that Ag_i can still get a trust value indirectly if it exists a path of trust in the trust network, going from Ag_i to Ag_j. Until now, this reasoning is fine by me.

They assume that this trust value is the minimal trust value in the path going from Ag_i to Ag_j. If there are multiple paths, they consider the maximal value among the minimal values returned by each path.
The problem here is that authors should introduce another kind of trust: the trust that the trustor has in who is reporting the trust values. First of all, I should understand how much the trustee that I am considering is motivated to give me the information and if it wants to deceive me. Even if I am sure that agent Y will correctly report its evaluation of Z to me, how much am I sure that Y is able to evaluate Z?
This second trust value is completely different from the trust evaluation of Y has a source as concern its ability to evaluate a source, not just to report an information. It is a completely different task. It should be used and taken into account for computing indirect trust.

This work allows the presence of multiple information sources, which have the possibility to assert different things. However they can just support o deny a specific belief, they don’t have other possibilities. Thanks to this simplification, authors can use a very simple computational model.
Moreover agents are classified into different categories, according to how they use their meta-information.

Interestingly authors take into consideration the notion of time. In fact if we are in a dynamic context, in which the state of the world can change continuously, it is better to get fresher information. This model allows the possibility to give priority to fresher information rather than the old one. However, it seems that the two level of trust and time are considered separately, there is no an explicit relationship between them.

An Argumentation-based Approach for Reasoning about Trust in Information Sources

In (Amgoud and Demolombe, 2014) authors propose an argumentation based system in which agents by the means of argumentation reason about their internal belief and the information received by their sources. According to this framework, agents will decide to believe or not in a given statement, how much a source is trustworthy and then to accept or not the belief that the source reports.

A good point of this work is the use of a fine-grained notion of trust. In fact this concept is based on six trustee's properties, already identified in (Demolombe 1999) (Demolombe, 2004): validity, completeness, sincerity, cooperativeness, competence and vigilance. For instance, an agent a_i can trust another agent a_j for its sincerity, but not for its competence.

The trustor will reason about them in order to evaluate the trustee, and then what it is reporting.

Arguing about the trustworthiness of the information sources

Even (Villata et al, 2011) propose to use argumentation to verify the validity of information. In this work authors suppose that the information can be attacked both by the means of argumentation and by the lack of trust. If I trust enough the information source in reporting me the correct information, then I will accept it as true. But the lack of trust implies that the information will not be accepted.

In fact each information is attacked by the default attack that checks if it is acceptable. Then in order to be accepted information needs to be supported by another argument, which is "source i is trustable". In other words, trust is a meta-argument used to attack this default auxiliary argument. The trust level can be enough to accept the information or not. For example there could be another argument attacking it. In this second case it is necessary to have other supporting evidence.

Then in this model "trust is represented by default as the absence of an attack towards the sources or towards the information items and as the presence of evidence in favor of the pieces of information. On the contrary, the distrust relationship is modeled as a lack of evidence in support of the information items or as a direct attack towards the sources and their pieces of information".

An Argumentation-Based Approach to Handling Trust in Distributed Decision Making

Another approach that exploits argumentation is the one proposed by (Parsons et al, 2013). In their work they want to understand how much information coming from a specific source should impact on the final decision in presence of multiple sources. In order to distinguish among sources, they use the concept of trust. In particular, they exploit the model of trust proposed by (Castelfranchi and Falcone, 2000) as it is based on reason, believing that "people are able to take critical decisions under time pressure soundly and with high confidence only if they understand the bases for their decisions".

In their approach they use formal argumentation, as it records the steps that leads to a conclusion, so it is really easy to exploit them for the reasoning process. Here each agent possesses a specific knowledge base and is related with the other agents through a social network indicating how much it trusts the agents it knows.

The platform takes as input an XML file, containing both the description of the trust network (agent X trust agent Y with a degree D) and the knowledge that each agent possesses. Even this data has e degree of belief, which represents how much the agent believes that its own information is true.

This is an interesting work, but I believe there is some lack concerning how the authors deal with trust. The relationship of trust between a trustor and a trustee is described by a trust value. This value is used as a weight to give more or less importance to information reported by a source. But there is no mention about how this value is used to aggregate what the source is reporting with information coming from other sources. This process is not explained in the paper.

Again, trust can be propagated in a trust network so that it is possible to compute it even for agents that are not related. How do the authors do it? I cannot find any explanation in the paper or a reference to other works.

Belief revision process based on trust: Simulation experiments

In (Barber and Kim, 2001a) authors apply the concepts of trust and reputation to information sources.
They start from the consideration that it is not possible for an individual agent to possess the global truth. Each of them will just possess a subjective perspective, and this clearly leads to a situation in which they have different information. Actually this is a reasonable choice in many contexts, as it is very difficult that just one agent possess all the knowledge (or as the same authors note, that it can afford the cost of getting and maintaining this knowledge).

Here trust is defined as the "confidence in the ability and intention of an information source to deliver correct information". According to that, they also propose a definition for reputation as "the amount of trust an information source has created for itself through interactions with other agents". Both these definitions are contextualized to the specific topic of information.

This work proposes an interesting architecture for modeling agents, based on modules that take care of specific functionalities. However the most interesting section, regarding the work presented in this thesis, is the belief revision process works. Let's see it in details.

Authors propose the division of agents' knowledge into two sets. The first one is called background knowledge base (KB) and includes all the knowledge that an agent accumulates in its life. Of course, because of its nature, it is can be (and often it is) inconsistent.

This first kind of memory differs from working knowledge base (K), that is the actual set of information that an agent use to reason and to generate behaviors. It is generated starting from the KB, as it is a maximally consistent set of knowledge derived from it. The consistency is a necessity, as agents need it for their reasoning process.

Agents' exchange of information is represented by the means of propositional language. For example $send(S_1,X,q,\alpha)$ means that the information source S_1 is sending to the agent X the information q and S_1 believes that q is true with a probability of α (that is a real number). This is an interesting point, as the source in not just reporting information but it adds as meta-information its personal degree of trust on its information. Actually, there would be a lot to say about that. X cannot be sure that S_1 truly believes that q is true with a probability of α. This presuppose that S_1 is sincere, that it has no reason to deceive me. It's faults could not just be related to its lack of competence; one should also consider the motivational point of view: is S_1 really interested to report me correct information?

Once X receives the information, it processes it according to its internal beliefs and it will consider q true with a probability α^1, which is based on the evidence α reported by S_1 but can be different.

To produce the certainty X has on q, it computes a polytree considering all the information coming from the source reporting q and their relative degree of reliability. The value of q is obtained propagating the probabilities through the tree.

It is necessary to do this procedure for all the beliefs in KB. After that, it is possible to produce K. When there is conflict between two beliefs, the one with the maximal value of q is kept, while the other is discarded. If the value is the same we take one randomly (because we want K to be maximal).

This model has been successfully used in another work of the same authors (Barber and Kim, 2001b) in which they show how agents are able to identify and eliminate unreliable information and unreliable source.

CHAPTER 2:

THE COMPUTATIONAL MODEL OF TRUST ON INFORMATION SOURCES AND ITS EVOLUTION

Once identified in the cognitive model of trust a good instrument for trust analysis, we decided to start from it to build a theory about trust on information sources. This work has been developed in different moment of time, improving it progressively.

Starting from a basic model that allowed just the presence of a single information source, but that showed a complete analysis of the cognitive variables affecting this kind of trust, the next step has been a model able to process multiple sources but expressively limited: each source had just the possibility to affirm or deny a specific belief. In the end, we developed a model that removed each of these limitations, exploiting the Bayesian theory.

In this chapter we are going to see all these models, showing how each of them contributed to the final result. It is worth noting that these models are not context dependent, so they can be basically used in a lot of domains.

The work presented in this chapter has resulted in the following publications:

1. Castelfranchi C., Falcone R., Sapienza A., Information sources: Trust and meta-trust dimensions. In proceeding of the workshop TRUST

2014, colocated with AAMAS 2014, Paris, CEUR Workshop Proceedings 2014 (In press)

2. Castelfranchi C., Falcone R., Sapienza A., Who to believe? Trust and meta-trust dimensions for Information sources. In proceedings of the conference Arguing on the Web 2.0, 2014 (in press)
3. Sapienza, A., Falcone, R., & Castelfranchi, C. “Trust on Information Sources: A theoretical and computation approach”, WOA 2014, Catania, Ceur-ws proceedings, Volume 1260, paper 12.
4. Sapienza, Alessandro, and Rino Falcone. "A Bayesian Computational Model for Trust on Information Sources.", WOA 2016, Catania, Ceur-ws proceedings, Volume 1664, pp. 50-55.

1. THE MODEL WITH ONE SOURCE

Introduction

According to the perspective of the cognitive model of trust (Castelfranchi et al, 2003) (Castelfranchi and Falcone, 2010) trust in information sources is just a kind of social trust, preserving all its prototypical properties and dimensions; just adding new important features and dynamics. In particular, also the trust in information sources (Demolombe, 1999) can just be an evaluation, judgment and feeling, or a decision to rely on, and the act of believing in and to the trustee (Y) and rely on it.

This trust and the perceived trustworthiness of Y has two main dimensions: the ascribed competence versus the ascribed willingness (intentions, persistence, reliability, honesty, sincerity, etc.). Also this form of trust is not empty, without a, more or less specified, object/argument: ”X trusts Y”! As Castelfranchi and Falcone shown, trust is for/about something, it has an object: what X expects from Y; Y’s service, action, provided good. And it is also context-dependent: in a given situation; with internal or external causal attribution in case of success or failure (trust in the agent versus trust in the environment). Also this form of trust is gradable: ”X trusts more or less Y”.

What changes is just the service and good X is expecting from Y, that is reliable knowledge. Thus all those dimensions are specified in this direction and acquire special qualities. For example, we can rely on Y without directly knowing it (and without specific recommendation from others) just because Y inherits the trust we have in a given class of people or group (if Y is assumed to be a good member, a typical instance of it: degree of) (Falcone et al, 2013). In a practical domain if I trust fire fighters or specialized plumbers I can trust Y’s practical ability as a fire-fighter, as a professional plumber; analogously, if I trust medical doctors’ competence, or press agencies, I will believe the recommendations of this doctor or the news of this press agency. Information is a resource, like others; more or less relevant (Paglieri and Castelfranchi, 2012) and good or bad for our goals. And providing relevant information is a ”service” like others.

In particular, in this work we are interested in the fact that the relevance and the trustworthiness of the information acquired by an agent X from a source F, strictly depends and derives from the X’s trust in F with respect the kind of that information.

Dimensions of Trust in Information Sources

Given the frame described above, which are the important specific dimensions of trust in information sources (TIS)? Many of these dimensions are quite sophisticated, given the importance of information for human activity and cooperation. We will simplify and put aside several of them.

First of all, we have to trust (more or less) the source (F) as competent and reliable in that domain, in the domain of the specific information content. Am I waiting for some advice on train schedule? On weather forecast? On the program for the examination? On a cooking recipe? Is this F not only competent but also reliable (in general or specifically towards me)? Is F sincere and honest? Is it leaning to lie and deceive? Will F do what has promised to do or ”has” to do for his role? And so on.

These competence and reliability evaluations can derive from different reasons, basically:

1. Direct experience with F (how F performed in the past interactions) on that specific information content;
2. Recommendations (other individuals Z reporting their direct experience and evaluation about F) or Reputation (the shared general opinion of others about F) on that specific information content (Yolum and Singh, 2003) (Conte and Paolucci, 2002) (Sabater-Mir, 2003) (Sabater-Mir and Sierra, 2001) (Jiang et al, 2013);
3. Categorization of F (it is assumed that a source can be categorized and that it is known this category): on this basis it is also possible to establish the competence/reliability of F on the specific information content (Falcone et al, 2013)(Burnett et al, 2010);

The two faces of F’s trustworthiness (competence and reliability) are relatively independent[1]; we will treat them as such. Moreover, we will simplify these complex components in just one quantitative fuzzy parameter by combining competence and reliability: F’s estimated trustworthiness. In particular we define the following fuzzy set: terrible, poor, mediocre, good, excellent (see figure 1) and apply it to each of the previous different dimensions (direct experience, recommendations and reputation, categorization).

Second, information sources have and give us a specific information that they know/believe; but believing something is not a yes/no status; we can be more or less convinced and sure (on the basis of our evidences, sources, reasoning). Therefore a good source might not just inform us not about P, but also about its degree of certainty about P, its trust in the truth of P. For example: ”It is absolutely sure that P”, ”Probably P”, ”It is frequent that P”, ”It might be that P”, and so on.

Of course there are more sophisticated meta-trust dimensions like: how much am I sure, confident, in F’s evaluation of the probability of the event or in his subjective certainty?[2] Is F not sincere? Is it enough self-confident and good evaluator? For example, in drug leaflet they say that a given possible bad side effect is only in 1% of cases. Have I to believe that? Or they are not reliable since they want to sell that drug? For the moment, we put aside that dimension of how much meta-trust we have in the provided

[1] Actually they are not fully independent. For example, F might be tempted to lie to me if/when is not so competent or providing good products: he has more motives for fudging me.

[2]In a sense it is a transitivity principle (Falcone and Castelfranchi, 2012): X trust Y, and Y trust Z; will X trust Z? Only if X trusts Y ”as a good evaluator of Z and of that domain”. Analogously here: will X trust Y because Y trusts Y? Only if X trust Y ”as a good and reliable evaluator” of it-self.

degree of credibility. We will just combine the provided certainty of P with the reliability of F as source. It in fact makes a difference if an excellent or a mediocre F says that the degree of certainty of P is 70%.

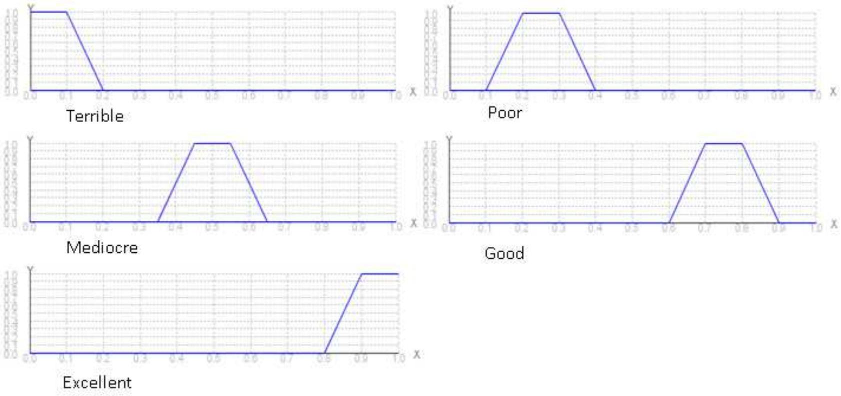

Figure 1: Representation of the five fuzzy sets

Third, especially for information sources it is very relevant the following form of trust: the trust we have that the information under analysis derives from that specific source, how much we are sure about that ”transmission”; that is, that the communication has been correct and working (and complete); that there are no interferences and alterations, and I received and understood correctly; that the F is really that F (Identity). Otherwise I cannot apply the first factor: F’s credibility. Let’s simplify also these dimensions, and formalize just the degree of trust that F is F; that the F of that information (I have to decide whether believe or not) is actually F. In the WEB this is an imperative problem: the problem of the real identity of the F, and of the reliability of the signs of that identity, and of the communication.

These dimensions of TIS are quite independent of each other (and we will treat them as such); we have just to combine them and provide the appropriate dynamics. For example, what happen if a given very reliable source F' says that "it is sure that P", but I'm not sure at all that the information really comes from F' and I cannot ascertain that?

Additional problems and dimensions

We believe in a given datum on the basis of its origin, its source: perception? communication? inference? And so on.

- The more reliable (trusted) the F the stronger the trust in P, the strength of the Belief that P. This is why it is very important to have a "memory" of the sources of our beliefs.

However, there is another fundamental principle of the degree of credibility of a given Belief (its trustworthiness):

- The many the converging sources, the stronger our belief (of course, if there are no correlations among the sources).

Thus we have the problem to combine different sources about P, their subjective degrees of certainty and their credibility, in order to weigh the credibility of P, and have an incentive due to a large convergence of sources. There might be different heuristics for dealing with contradictory information and sources. One (prudent) agent might adopt as assumption the worst hypothesis, the weaker degree of P; another (optimistic) agent, might choose the best, more favorable estimation; another agent might choose the most reliable source. We will formalize only one strategy: the weighing up and combination of the different strengths of the different sources, avoiding however the psychologically incorrect result of probability values, where by combining different probabilities we always decrease the certainty, it never increases. On the contrary - as we said - convergent sources reinforce each other and make us more certain of that datum.

Feedback on source credibility/TIS

We have to store the sources of our beliefs because, since we believe on the basis of source credibility, we have to be in condition to adjust such credibility, our TIS, on the basis of the result. If I believe that P on the basis of source F1, and later I discover that P is false, that F1 was wrong or deceptive, I have to readjust my trust in F1, in order next time (or with similar sources) to be more prudent. And the same also in case of positive confirmation.[3]

However, remember that it is well known (Urbano et al, 2009) that the negative feedback (invalidation of TIS) is more effective and heavy than the positive one (confirmation). This asymmetry (the collapse of trust in case on negative experience versus the slow acquisition or increasing of trust) is not specific of trust and of TIS; it is -in our view- basically an effect of a general cognitive phenomenon. It is not an accident or weirdness if the disappointment of trust has a much stronger (negative) impact than the (positive) impact of confirmation. It is just a sub-case of the general and fundamental asymmetry of negative vs. positive results, and more precisely of ”losses” against ”winnings”: the well-known Prospect theory (Kahneman and Tversky, 1979). We do not evaluate in a symmetric way and on the basis of an ”objective” value/quantity our progresses and acquisitions versus our failures and wastes, relatively to our ”status quo”. Losses (with the same ”objective” value) are perceived and treated as much more severe: the curve of losses is convex and steep while that of winnings is concave. Analogously the urgency and pressure of the ”avoidance” goals is greater than the impulse/strength of the achievement goals (Higgins, 1997). All this

[3] We can even memorize something that we reject. We do not believe to it but not necessarily we delete/forget it, and its source. This is for the same function: in case that information would result correct I have to revise my lack of trust in that source.

applies also to the slow increasing of trust and its fast decreasing; and to the subjective impact of trust disappointment (betrayal!) vs. trust confirmation. That's why usually we are prudent in deciding to trust somebody; in order do not expose us to disappointment and betrayals, and harms. However, also this is not always true; we have quite naive forms of trust just based on gregariousness and imitation, on sympathy and feelings, on the diffuse trust in that environment and group, etc. This also plays a crucial role in social networks on the web, in web recommendations, etc.

Moreover, according to (Falcone and Castelfranchi, 2004) a bad result (or a good result) not always and automatically entails the revision of TIS. It depends on the ”causal attribution”: has it been a fault/defect of F or interference on the environment? The result might be bad although F's performance was his best. Let us put aside here the feedback effect and revision on TIS.

Where does TIS comes from?

Which are the sources, bases, of our trust in our information sources? What does determine our trust in their message/content? Let's add something more specific on what we have already seen, also because in the literature ”direct experience” and reputation-recommendation play an absolutely dominant role:

1. Our previous direct experience with F, or better our ”memory” about, and the adjustment that we have made of our evaluation of F in several interaction, and possible successes or failure relying on its information.
2. The others' evaluation/trust; either inferred from their behavior or attitude by ”transitivity”: ”If Y trusts Z, me too can trust Z”, or due to explicit recommendation from Y (and others) about Z; or due to Z's

”reputation” in that social environment: circulating, emergent opinion.

3. Inference and reasoning:
    a. by inheritance from classes or groups were Z id belonging (as a good ”exemplar”);
    b. by analogy: Z is (as for that) like Y, Y is good for, then Z too is good for;
    c. by analogy on the task: Z is good/reliable for P he should be good also for P’, since P and P’ are very similar. (In any case: how much do I trust my reasoning ability?).

Plausibility: the integration with previous knowledge

To believe something means not just to put it into a file in my mind; it means to ”integrate” it with my previous knowledge. Knowledge must be at least non-contradictory and possibly supported, justified: this explains that, and it is explained, supported, by these other facts/arguments. If there is contradiction I cannot believe P; either I have to reject P or I have to revise my previous beliefs in order to coherently introduce P. It depends on the strength of the new information (its credibility, due to its sources) and on the number and strength of the internal opposition: the value of the contradictory previous beliefs, and the extension and cost of the required revision. That is: it is not enough that the candidate belief that P be well supported and highly credible; is there an epistemic conflict? Is it ”implausible” to me? Are there antagonistic beliefs? And which is their strength? The winner of the conflict will be the stronger ”group” of beliefs. Even the information of a very credible source (like our own eyes) can be rejected!

Trusting as Risking

In which sense and on what ground do we make ”reliance” on a belief? We decide and pursue a given goal precisely on the basis of what we believe, and of our certainty on current circumstances and predictions. We reasonably invest and persist in a given activity proportionally to our degree of certainty, of confidence: the more we believe in it the more we bet on and take risks (we sacrifice alternatives and resources, and expose ourselves to possible failures, wastes, and regrets). In other words, our trust in a given information or fact exposes us to ”risks”. Trust always means to take same risk. As for TIS our claim is that:

> - *the higher the perceived risk* (estimated probability and gravity of the failure plus probability and gravity of the possible harms) *the higher the threshold for accepting a given belief and for deciding (trust as decision) to act on such a basis*.

In this case, we would search or wait for additional data and sources. We will not formalize in this work this crucial aspect.

Layered trust

Notice the recursive and inheritance structure of trust and the various meta-levels: I rely and risk on a given assumption on the basis of how much I believe in it, but I believe in it (trust in Belief) on the basis of the number, convergence, and credibility of the information source (TIS) and on the trust they have in what they ”say” (and on the trust I have in the trust they have on what they say: meta-meta-trust).

I trust a given source on the basis of information about it: from my memory and experience, from inference, from others (transitivity, reputation, or recommendation), then I have to trust these new information sources: the information about my information sources. And so on (Castelfranchi and Falcone, 2010).

Formalizing and computing the degree of certainty as trust in the belief

As we already said, there is a confidence, a trust in the beliefs we have and on which we rely. Suppose X is a cognitive agent, an agent who has beliefs and goals. Given Bel_X, the set of the X's beliefs, then P is a belief of X if:

$P \in Bel_X$ (1)

The degree of subjective certainty or strength of the X's belief P corresponds with the X's trust about P, and call it:

$Trust_X (P)$ (2)

Its origin/ground

In the case in which we are considering just specific information, P, deriving from just one source F, $Trust_X(P)$ depends on:

i) the X's trust towards F as source of the information P (that could mean with respect the class of information to which P belongs):
$Trust_X (F,P)$ (3)
and

ii) the relationships between P and the other X's beliefs.

With respect to the case (ii), if we term Q a X's belief $Q \in Bel_X$ and $Trust_X(P|Q)$ the X's Trust in P given the X's belief Q; we can say that:

1. if $Trust_X(P|Q) > Trust_X(P)$ (4)
   Q positively interferes with P (Q supports P);
2. if $Trust_X(P|Q) < Trust_X(P)$ (5)

Q negatively interferes with P (Q is in contradiction with P).

So we can say that given a X's belief Q, it can have positive, negative or neutral interference with the new information P that X is acquiring. The value of this interference (in the positive or negative cases) will be called

$\Delta INT_{X,P}$ (Q) :

$\Delta INT_{X,P}$ (Q) = $|Trust_X(P|Q) - Trust_X(P)|$ (6)

It is a function of two main factors, the X's trust in Q and the strength of the Q's interference with P:

$\Delta INT_{X,P}$ (Q) = $f_1(Trust_X(Q),DInt(Q,P))$ (7)

It is important to underline that we are considering the composition among different information (P, Q_i) just from a very abstract and not analytical point of view. In this work we do not cope with the relevant problems studied in this domain by the research on argumentation (Walton, 1996)(Walton et al, 2008).

Then supposing there is just one of X's belief (Q) interfering with P, we have that:

$Trust_X(P) = f_2(Trust_X(F,P),\Delta INT_{X,P}$ (Q)) (8)

In other words: the X's trust in information P is a function of both the X's trust in the source F about P and of the interference factor between P and Q (as showed by $\Delta INT_{X,P}$ (Q)). In the case of more than one X's beliefs interfering with P, say Q_i (with i= 1, ..., n), we have to compose the n interfering factors ($\Delta INT_{X,P}$ (Q_i)) in just one resulting factor.

Applying now the conceptual modeling previously described in the we have that $Trust_X(F,P)$ can be articulated in:

1. the X's trust about P just depending from the X's judgment of the F's competence and reliability as derived from the composition of the three factors (direct experience, recommendation/reputation, and categorization), in practice the F's credibility about P on view of X:

   $Trust^1_X(F,P)$ (9)
2. the F's degree of certainty about P: information sources give not only the information but also their certainty about this information; given that we are interested to this certainty, but we have to consider that through X's point of view, we introduce

   $Trust_X(Trust_F (P))$ (10)

   In particular, we consider that X completely trusts F, so that $Trust_X(Trust_F (P)) = Trust_F(P)$
3. the X's degree of trust that P derives from F: the trust we have that the information under analysis derives from that specific source:

   $Trust_X(Source(F, P))$ (11)

Resuming:

$Trust_X (F, P) = f3(Trust^1_X(F, P), Trust_X(Trust_F(P)), Trust_X(Source(F,P)))$ (12)

Here we could introduce a threshold for each of these 3 dimensions, allowing reducing risk factors.

A modality of computation

$Trust^1_X$ (F, P)

As previously specified, the value of $Trust^1_X(F,P)$ is a function of:

1. Past interactions;
2. The category of membership;
3. Reputation.

Each of these values is represented by a fuzzy set: terrible, poor, mediocre, good, excellent. We then compose them into a single fuzzy set, considering a weight for each of these three parameters. Those weights are defined in range [0;10], with 0 meaning that the element has no importance in the evaluation and 10 meaning that it has the maximal importance. It is worth noting that the weight of experience has to be referred to a twofold meaning: it must take into account the numerosity of experiences (with their positive and negative values), but also the intrinsic value of experience for that subject.

However, the fuzzy set in and by itself is not very useful: what interests us in the end is to have a plausibility range, which is representative of the expected value of $\text{Trust}^1{}_X(F,P)$. In order to get that, it is therefore necessary to apply a defuzzyfication method. Among the various possibilities (mean of maxima, mean of centers) we have chosen to use the centroid method, as we believed it can provide a good representation of the fuzzy set. The centroid method exploits the following formula:

$$k = \quad (\int_0^1 x\ f(x)\ dx) / \ (\int_0^1 f(x)\ dx)$$

were f(x) is the fuzzy set function. The value k, obtained in output, is equal to the abscissa of the gravity center of the fuzzy set. This value is also associated with the variance, obtained by the formula:

$$\sigma^2 = (\int_0^1 (x - k)^2\ f(x)\ dx) / \ (\int_0^1 f(x)\ dx)$$

With these two values, we determine $\text{Trust}^1{}_X(F,P)$ as the interval $[k - \sigma;\ k +$

σ].

TrustX(F,P)

Once we get $\text{Trust}^1{}_X$(F,P), we can determine the value of Trust_X(F,P). In particular, we determine a trust value followed by an interval, namely the uncertainty on Trust_X(F,P). For uncertainty calculation we use the formula:

$\text{Uncertainty} = 1 - (1 - \Delta x) * \text{Trust}_X(\text{Trust}_F\,(P)) * \text{Trust}_X(\text{Source}(F,P))$

$\Delta x = \text{Max}(\text{Trust}^1{}_X\,(F,P)) - \text{Min}(\text{Trust}^1{}_X(F,P))$

In other words, the uncertainty depends on the uncertainty interval of $\text{Trust}^1{}_X$(F,P), properly modulated by Trust_X (Trust_F(P)) and Trust_X(Source(F,P)). This formula implies that uncertainty:

- Increase/decrease linearly when Δx increase/decrease;
- Increase/decrease linearly when $\text{Trust}_X(\text{Trust}_F(P))$ decrease/increase;
- Increase/decrease linearly when $\text{Trust}_X(\text{Source}(F,P))$ decrease/increase;

The inverse behavior of $\text{Trust}_X(\text{Trust}_F(P))$ and $\text{Trust}_X(\text{Source}(F,P))$ is perfectly explained by the fact that when X is not so sure that P derives from F or F's degree of certainty about P is low, global uncertainty should increase. The maximum uncertainty value is 1 (+-50%) meaning that X is absolutely not sure about its evaluation. On the contrary, the minimum value of uncertainty is 0, meaning that X is absolutely sure about its evaluation.

In a way similar to uncertainty, we used the following formula to compute a value of $\mathrm{Trust}_X(F,P)$:

$\mathrm{Trust}_X(F,P) = 1/2 + (\mathrm{Trust}^1_X(F,P) - 1/2) * \mathrm{Trust}_X(\mathrm{Trust}_F(P)) * \mathrm{Trust}_X(\mathrm{Source}(F,P))$

This formula has a particular trend, different from that of uncertainty. Here in fact the point of convergence is 1/2, value that does not give any information about how much X can trust F about P. Notice that:

1. If $\mathrm{Trust}^1_X(F,P)$ is less than 1/2, as $\mathrm{Trust}_X(\mathrm{Trust}_F(P))$ and $\mathrm{Trust}_X(\mathrm{Source}(F,P))$ increase the value of trust will decrease going to the value of $\mathrm{Trust}^1_X(F,P)$; as $\mathrm{Trust}_X(\mathrm{Trust}_F(P))$ and $\mathrm{Trust}_X(\mathrm{Source}(F,P))$ decrease the value of trust will increase going to 1/2;
2. If $\mathrm{Trust}^1_X(F,P)$ is more than 1/2, as $\mathrm{Trust}_X(\mathrm{Trust}_F(P))$ and $\mathrm{Trust}_X(\mathrm{Source}(F,P))$ increase the value of trust will increase going to the value of $\mathrm{Trust}^1_X(F,P)$; as $\mathrm{Trust}_X(\mathrm{Trust}_F(P))$ and $\mathrm{Trust}_X(\mathrm{Source}(F,P))$ decrease the value of trust will decrease going to 1/2.

Computing a final trust value

Here we use an aggregation formula based on the sinusoidal function (Urbano et al, 2009):

$\gamma = \delta * \sin(\alpha) + \delta$

in which:

$\delta = 1/2$;

$\omega_0 = \pi;$

$\omega = \omega_0/8;$

$\alpha_0 = 3/2\pi;$

$\alpha = \alpha_0 + \omega_0 * Trust_X(F,P) + PP * \omega * (\Sigma\Delta Supp\ INT_{X,P}\ (Q)) + PN * \omega * (\Sigma\Delta Opp\ INT_{X,P}\ (Q))$

ω is the aggregation step. Increasing this value, a high number of supporting/opposing beliefs will easily let the trust value converge to 1 or 0 (it depends from the division factor: we chose 8). PP and PN are respectively the weights of positive and negative beliefs. Of course, they are function of the subjective value of the goal and the risk that the trustor is willing to address. However, in this work we assume them to be constant.

PP = 1
PN = 1

Let us specify a methodological aspect. The choice of the parameters of the previous formulas should emerge from the many features of the domain (types of agents, information, sources, and so on) and from experiments simulating them. In fact we are introducing these formulas just considering a top-down model, so our choices of these parameters are quite arbitrary.

How previously stated, interference of a single belief is determined by X's trust in Q and the strength of the Q's interference with P. In particular

$\Delta INT_{X,P}\ (Q) = Trust_X(Q) * DInt(Q,P)$

DInt(Q,P) is defined it the range [-1,1], with a value of -1 meaning totally opposing and 1 totally supporting. Moreover each interference (actually its absolute value) is compared to a threshold, before being taken into account in the final calculation: if the value is less than the threshold it is discarded.

Some interesting examples

Let us now present some examples of interest.

EXAMPLE 1

In the first example we show hot to compute a value of Trust$^{1}_{X}$(F,P). In this particular situation, X gives really more importance to past experience (maximum weight value) than to category (medium weight value). Here reputation is not so important.

Table 1: INPUT DATA

Reputation	good
Reputation weight	2
Category	mediocre
Category weight	5
Past Experience	excellent
Past Experience weight	10

Table 2: OUTPUT DATA

Mean	0.7511
Variance	0.04119

The result is that we have a quite high value of Trust$^{1}_{X}$(F,P), with low variance.

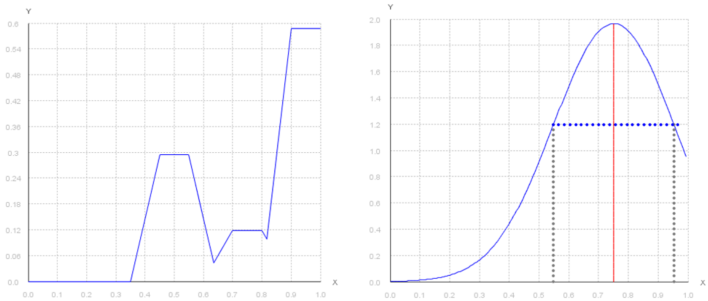

Figure 2: Global Fuzzy set and Gaussian Trend in the example

EXAMPLE 2

In a second situation, X has not past experience to use in his evaluation, but it can still assess a value of $\text{Trust}^1{}_X(F,P)$, basing its evaluation on reputation and category. Supposing that it is in an environment in which there is a high reputation, it decides to give more importance to this factor.

Table 3: INPUT DATA

Reputation	excellent
Reputation weight	10
Category	mediocre
Category weight	2
Past Experience	none
Past Experience weight	-

Table 4: OUTPUT DATA

Mean	0.83333
Variance	0.0323

Here are the results:

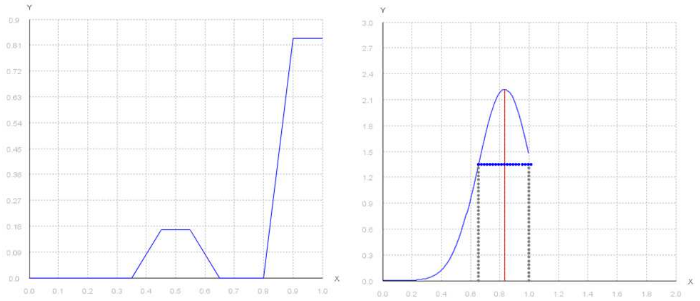

Figure 3: Global Fuzzy set and Gaussian Trend in the example

EXAMPLE 2.1

Now, fixing the value of Trust$^1{}_X$(F,P) from the last example, let's see how Trust$_X$(Trust$_F$(P)) and Trust$_X$(Source(F,P)) influence the value of Trust$_X$(F,P).

Table 5: INPUT 1

Trust$_X$ (Trust$_F$(P))	1
Trust$_X$(Source(F,P))	1

Table 6: OUTPUT 1

Trust$_X$(F,P)	0.83333
Uncertainty	+-17.32%

In this case, the source F is sure about the belief P, then Trust$_X$(Trust$_F$(P)) = 1. Moreover X is sure that the source of P is exactly F, then Trust$_X$(Source(F,P)) = 1. This implies that Trust$_X$(F,P) is exactly equal to Trust$^1{}_X$(F,P).

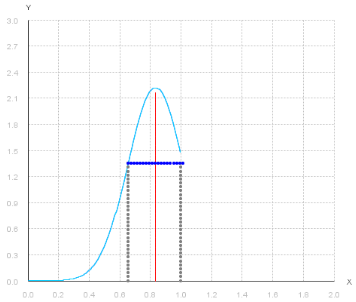

Figure 4: Gaussian Trend

EXAMPLE 2.2

Here X evaluates an uncertainty of F about the belief P.

Table 7: INPUT 2

Trust$_X$(Trust$_F$(P))	0.5
Trust$_X$(Source(F,P))	1

Table 8: OUTPUT 2

Trust$_X$(F,P)	0.66666
Uncertainty	+-33.66%

We can see a tendency of Trust$_X$(F,P) to decrease (such decrease has as a lower limit the maximal trust ambiguity, that is 0.5) and uncertainty increases towards its maximal value, that is +-0.5.

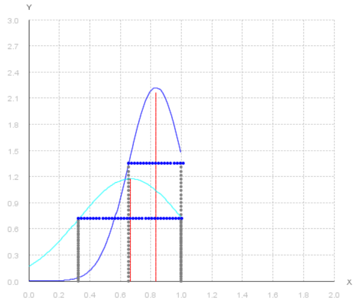

Figure 5: Gaussian Trend

EXAMPLE 2.3

Table 9: INPUT 3

$Trust_X(Trust_F(P))$	0.5
$Trust_X(Source(F,P))$	0

Table 10: OUTPUT 3

$Trust_X(F,P)$	0.5
Uncertainty	+-50.0%

Here X is in the worst situation, it cannot assume anything. This makes perfect logical sense because it is sure F is not the source of P.

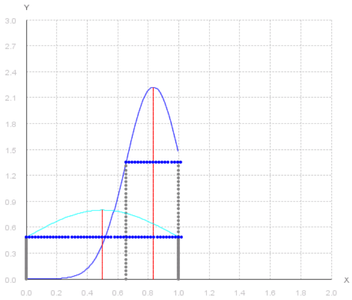

Figure 6: Gaussian Trend

Conclusions

Let us conclude with the ”Liar Paradox”: a specific distrust in F’s sincerity and honesty, that is the strong belief that F is lying while asserting that P, has a paradoxical consequence: *I can reasonably come to believe the opposite*. If F is lying and thus P is false, it is true (I have to believe) that Not(P).

This does not necessarily mean that I know how the world is. If there is a two-value world/case (P or Q), given that P is false I'm sure that Q (I can trust the belief that Q given my distrust in F!). But, if the case/world has multiple alternative (P or Q or W or Z), I can just believe (with certainty) that given Not (P) it is either Q or W or Z. Two burglars are trying to break a shop open when the police shows up. One burglar manages to fly away, but the other is caught by the policemen, whose ask him which way his accomplice went. ”That way!”, answers the burglar, while pointing to the right. And the policemen run in the opposite direction!

Notice that when my distrust is about the *competence* dimension, that is I'm sure that F is not expert and informed at all about P, F's assertion that ”surely P” doesn't give me the certainty that Not (P), but just leave me fully *uncertain*: I don't know, I cannot use F as source for P.

This different effect (especially for TIS) between the *competence dimension* of trustworthiness and the *honesty/sincerity* (reliability) dimension is quite interesting and can help to clarify the necessity of a rich analysis of the trust in information sources.

2. THE MODEL WITH MULTIPLE SOURCE ASSERTING (OR DENYING) A SINGLE BELIEF

Given that with the theoretical analysis of the previous section of this chapter, all the cognitive ingredients affecting trust on information sources have already been identified and studied, here we will just focus on how these concepts have been formalized into the computational model.

Formalizing and computing the degree of certainty as trust in the belief

As we already said, there is a confidence, a trust in the beliefs we have and on which we rely.

Suppose X is a cognitive agent, an agent who has beliefs and goals. Given Bel_X, the set of the X's beliefs, then P is a belief of X if:

$$P \in Bel_X \qquad (1)$$

The degree of subjective certainty or strength of the X's belief P corresponds with the X's trust about P, and call it:

$$Trust_X(P) \qquad (2)$$

Its origin/ground

Concerning a single belief P, we have to consider n different sources asserting or denying P. The final value of $Trust_X(P)$ depends on X's trust towards every single source F of the information P (that could mean with respect the class of information to which P belongs):

$$Trust_X(F,P) \qquad (3)$$

In other words, we state that:

$$\mathrm{Trust}_X(P) = f(\mathrm{Trust}_X(F_1,P), \ldots, \mathrm{Trust}_X(F_n,P)) \qquad (4)$$

Where n is the total number of sources.

Then to compute X's trust value, we have to compose the n sources' value in just one resulting factor.

Applying now the conceptual modeling previously described we have that $\mathrm{Trust}_X(F,P)$ can be articulated in:

1. X's trust about P just depending from the X's judgment of the F's competence and reliability as derived from the composition of the three factors (direct experience, recommendation/reputation, and categorization), in practice the F's credibility about P on view of X:
   $$\mathrm{Trust}^1{}_X(F,P) \qquad (5)$$
2. F's degree of certainty about P: information sources give not only the information but also their certainty about this information; given that we are interested to this certainty, but we have to consider that through X's point of view, we introduce
   $$\mathrm{Trust}_X(\mathrm{Trust}_F(P)) \qquad (6)$$
   in particular, we consider that X completely trusts F, so that $\mathrm{Trust}_X(\mathrm{Trust}_F(P)) = \mathrm{Trust}_F(P)$
3. the X's degree of trust that P derives from F: the trust we have that the information under analysis derives from that specific source:
   $$\mathrm{Trust}_X(\mathrm{Source}(F,P)) \qquad (7)$$
4. the fact that F is supporting P or is opposing to it (not P):
   $$\mathrm{Support}_F(P) \qquad (8)$$

Resuming:

$$\mathrm{Trust}_X(F,P) = f_3(\mathrm{Trust}^1{}_X(F,P), \mathrm{Trust}_X(\mathrm{Trust}_F(P)), \mathrm{Trust}_X(\mathrm{Source}(F,P)), \mathrm{Support}_F(P)) \qquad (9)$$

Here we could introduce a threshold for each of these 3 dimensions, allowing reducing risk factors.

A modality of computation

Trust^1_X(F,P)

As already specified, the value of Trust^1_X(F,P) is a function of:

1. Past interactions;
2. The category of membership;
3. Reputation.

Again each of these values is represented by a fuzzy set: terrible, poor, mediocre, good, excellent. We then compose them into a single fuzzy set, considering a weight for each of these three parameters. Those weights are defined in range [0;10], with 0 meaning that the element has no importance in the evaluation and 10 meaning that it has the maximal importance.

It is worth noting that the weight of the experience has to be referred to a twofold meaning: it must take into account the number of experiences (with their positive and negative values), but also the intrinsic value of experience for that subject.

However, the fuzzy set in and by itself is not very useful: what interests us in the end is to have a plausibility range, which is representative of the expected value of Trust^1_X(F,P).

To get that, it is therefore necessary to apply a defuzzyfication method. Among the various possibilities (mean of maxima, mean of centers ...) we have chosen to use the *centroid method*, as we believed it can provide a good representation of the fuzzy set. The centroid method exploits the following formula:

$$k = \left(\int_0^1 x\, f(x)\, dx\right) / \left(\int_0^1 f(x)\, dx\right) \qquad (10)$$

were f(x) is the fuzzy set function.

The value k, obtained in output, is equal to the abscissa of the gravity center of the fuzzy set.

This value is also associated with the variance, obtained by the formula:

$$\sigma^2 = (\int_0^1 (x - k)^2 f(x)\, dx)/ (\int_0^1 f(x)\, dx) \qquad (11)$$

With these two values, we determine $Trust^1_X(F,P)$. as the interval $[k-\sigma;k+\sigma]$.

$Trust_X(F,P)$

Once we get $Trust^1_X(F,P)$., we can determine the value of $Trust_X(F,P)$. In particular, we determine a trust value followed by an interval, namely the uncertainty on $Trust_X(F,P)$.

For uncertainty calculation we use the formula:

Uncertainty = $1 - (1- \Delta Trust^1_X)* Trust_X(Trust_F(P))* Trust_X(Source(F,P))$ *(12)*

$\boldsymbol{\Delta} Trust^1_X = Max(Trust^1_X(F,P)) - Min(Trust^1_X(F,P))$

In other words, the uncertainty depended on the uncertainty interval of $Trust^1_X(F,P)$, properly modulated by $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$.

This formula implies that uncertainty:

- Increase / decrease linearly when $\boldsymbol{\Delta} Trust^1_X$ increase / decrease;
- Increase / decrease linearly when $Trust_X(Trust_F(P))$ decrease / increase;

- Increase / decrease linearly when $Trust_X(Source(F,P))$ decrease / increase.

The inverse behavior of $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ is perfectly explained by the fact that, when X is not so sure that P derives from F or F's degree of certainty about P is low, global uncertainty should increase.

The maximum uncertainty value is 1 (+-50%) meaning that X is absolutely not sure about its evaluation. On the contrary, the minimum value of uncertainty is 0, meaning that X is absolutely sure about its evaluation.

In a way similar to uncertainty, we used the following formula to compute a value of $Trust_X(F,P)$:

1) If $Support_F(P) = 1$, namely F is supporting P

$Trust_X(F,P) = ½ + (Trust^1_X(F,P) – ½) * Trust_X(Trust_F(P)) * Trust_X(Source(F,P))$ (13a)

2) If $Support_F(P) = 1$, namely F is opposing P

$Trust_X(F,P) = ½ - (Trust^1_X(F,P) – ½) * Trust_X(Trust_F(P)) * Trust_X(Source(F,P))$ (13b)

This formula has a particular trend, different from that of uncertainty. Here in fact the point of convergence is ½, value that does not give any information about how much X can trust F about P. Notice that, if F is supporting P:

- If $Trust^1_X(F,P)$ is less than ½, as $Trust_X(Trust_F(P))$ and $Trust_X(Source(F,P))$ increase the value of trust will decrease going to the value of $Trust^1_X(F,P)$; as $Trust_X(Trust_F(P))$ and

$\text{Trust}_X(\text{Source}(F,P))$ decrease the value of trust will increase going to ½;

- If $\text{Trust}^1_X(F,P)$ is more than ½, as $\text{Trust}_X(\text{Trust}_F(P))$ and $\text{Trust}_X(\text{Source}(F,P))$ increase the value of trust will increase going to the value of $\text{Trust}^1_X(F,P)$; as $\text{Trust}_X(\text{Trust}_F(P))$ and $\text{Trust}_X(\text{Source}(F,P))$decrease the value of trust will decrease going to ½;

Conversely, when F is opposing P:

- If $\text{Trust}^1_X(F,P)$ is less than ½, as $\text{Trust}_X(\text{Trust}_F(P))$ and $\text{Trust}_X(\text{Source}(F,P))$ increase the value of trust will increase going to the value of $\text{Trust}^1_X(F,P)$; as $\text{Trust}_X(\text{Trust}_F(P))$ and $\text{Trust}_X(\text{Source}(F,P))$ decrease the value of trust will decrease going to ½;
- If $\text{Trust}^1_X(F,P)$ is more than ½, as $\text{Trust}_X(\text{Trust}_F(P))$ and $\text{Trust}_X(\text{Source}(F,P))$ increase the value of trust will decrease going to the value of $\text{Trust}^1_X(F,P)$; as $\text{Trust}_X(\text{Trust}_F(P))$ and $\text{Trust}_X(\text{Source}(F,P))$ decrease the value of trust will increase going to ½;

Computing a final trust value: sources' aggregation

How to evaluate the contribution of different sources? In general, the average value is given by the average of individual sources' trust value.

This issue gets more complicated when you need to find an average uncertainty value: computing the average of uncertainties is not enough. For instance, suppose we have two sources, the former asserting 0 with uncertainty 0 and the latter asserting 1 with uncertainty 0. Intuitively, a trust value of 0.5 is fine by me, but it is implausible that uncertainty is equal to 0; on the contrary, it should take the maximum value.

Thus it is easy to note how global uncertainty depends on both the single values of uncertainty and the single trust values. Plus we state that **the greater the number of convergent sources towards a trust value, the lower the uncertainty I have**. Then the formula to compute this global value should take into account these factors.

The domain of uncertainty [0,1] has been divided into 5 intervals of amplitude 0.2. Values falling in the same interval are considered convergent. Here is the used formula:

$$Unc = Unc_0 + \sum_j \sum_i^I Unc_i / (I*N) \qquad (14)$$

In which:

Unc_0 = minimum distance value between the computed medium trust value and each single trust value (of every single source);

j = intervals, 1<j<5;

I = number of convergent sources in the **j-th** interval;

N= total sources' number;

Unc_i = uncertainty on **i-th** source.

Thus it is worth noting that it is better to have two sources asserting the same thing, even if with a given value of uncertainty, than two sources asserting opposing information, even if with the utmost certainty.

Conclusion

In this work we have analyzed the nature of trust in information source also on the basis of our previous work (Castelfranchi et al, 2014).

We identified which components influence this kind of trust and showed how them contribute to the creation of trust. We also showed how the degree of trust in an information P strictly depends and derives from the X's trust in the sources producing it with respect the kind of information.

Finally we provided a detailed framework and a computational model to deal with this kind of problem.

We consider necessary to specify that, although we described the model and the variable that influence it, we have not investigated some important parameters (such as the weights of past experience, category and reputation). In fact we think that these values are strongly linked to the context in which the model is applied and should emerge from it.

3. THE MODEL WITH MULTIPLE SOURCE REPORTING DIFFERENT OPTIONS

This is the last version of the cognitive model we realized. It is very different from the other two, from a computational point of view as, in order to gain much more expressive power, we decided to exploit the Bayesian theory. In fact, thinking about the real world, it is just not binary (true or false) but there are a lot of possibilities. In this way, each source can express its subjective degree of belief that each possibility is true.

Introduction

In the world we often have to deal with different information coming from different information sources. Though having a lot of sources can be very useful, on the other hand trying to put together information coming from different information sources can be an uneasy task. It is necessary to have strategies to do it, especially in presence of critical situation, when there are temporal limits to make decision and a wrong choice can lead to an economical loss or even to risk life.

As said, the possibility of integrating sources on different scopes can be very useful in order to make a well-informed decision.
Integrating these sources becomes essential, but at the same time it is necessary to identify and take into account their trustworthiness.

Considering information's output, it can be just a true/false value (the source can just assert or deny the belief P) or there can be multiple outcomes. As this is a general model, we suppose that there can be different outcomes. For instance, the weather is not just good or bad, but can assume multiple values (critical, sunny, cloudy etc.).

The Bayesian Choice

There are many ways to computationally realize a decision making process and quite all of them provide good results.

Dealing with uncertain situations, one can use the uncertainty theory (Liu, 2014), a mathematical approach specifically created to evaluate belief degree in cases in which there is no data.

Another possible way is to use fuzzy logic (Sapienza et al, 2014). This technique has several vantages like:

1. It is flexible and easy to use;
2. It don't need precise data;
3. It can deal with non linear functions;
4. It is able to shape human way of think and express, as it can model concept that are more complex than a Boolean but not so precise like a real number.

Maybe the most used approach is the probabilistic one, which exploits the Bayesian theory, in particular probability distribution.

One of the advantages of using Bayesian theory is that it implies a sequential process: every time that new evidence occurs it can be processed individually and then aggregated to global evidence. This property is really useful as it allows a trustor to elaborate its information in a moment and update it whenever it gets other evidence.

Given the context of information sources, we believe that this last option is the choice that best suits with the problem. In fact there is a fixed number of known possibilities to model and the trustor can collect information from its sources individually and then aggregate them in different moment.
Plus, the scientific literature confirms its utility in the context of trust evaluation (Melaye and Demazeau, 2005) (Quercia et al, 2006) (Wang and Vassileva, 2003).

The Computation Model

In the proposed model each information source S is represented by a trust degree called $TrustOnS$, with 0≤ $TrustOnS$ ≤1, plus a bayesian probability distribution PDF that represents the information reported by S.

To the aim of granting a better flexibility, the PDF is modeled as a continuous distribution (actually it is divided into several intervals and it is continuous in each interval). In fact if the event domain is continuous it is better to use a continuous PDF; if it happens to be discrete it is still possible to use a continuous PDF. It is also possible to specify what and how much outcomes the model has to use, depending on the specific context. In the end of the chapter we will show a working example in which we take into account five different outcomes, then the PDF will be divided accordingly.

The model we created starts from a preliminary evaluation of the source trustworthiness: how much reliable is a source S concerning a specific information's category?
Then after evaluating it, we consider what the source is reporting - the PDF. We use the trust evaluation to understand how much the specific information should be considered, with respect to the global information.

This process can be done in presence of a single or multiple sources, as each time we perform an aggregation of each contribute to the global evidence.

A strong point of this model is that it is sequential, so it can be updated when new information comes.

Source's Evaluation

The first part of the model concerns the source's evaluation. According to us, there are two level of evaluation. Initially, we produce an a priori trust, which represent how much I believe that S is good with this specific kind of information.

After that, we compute a more sophisticated analysis taking into account other parameters.

Let's first start from the a priori source's evaluation – *SEvaluation*. **This is the trustor's trust about P just depending on the its judgment of the S's competence and willingness** as derived from the composition of the three factors (direct experience, recommendation/reputation, and categorization), in practice the S's credibility about P on view of the trustor. Recalling that a trust evaluation for a cognitive agent is based on the two aspects of competence and willingness, we state that these values can be obtained using three different dimensions:

1. **Direct experience** with S (how S performed in the past interactions) on that specific information content;
2. **Recommendations** (other individuals Z reporting their direct experience and evaluation about S) or **Reputation** (the shared

general opinion of others about S) on that specific information content;

3. **Categorization** of S.

The two faces of S's trustworthiness (competence and willingness) are relatively independent; however, for sake of simplicity, we will unify them into a unique quantitative parameter, by combining competence and reliability.

Computationally, the past experience (PE), reputation/recommendation (REP) and categories (CAT) parameters are defined here as real values in the interval [0,1]. To compute S's evaluation we make a weighted mean of them:

$$SEvaluation = w1 * PE + w2 * CAT + w3 * REP$$

The trustor, considering both its personality and the context in which it is, determines the weight w1, w2 and w3 empirically.

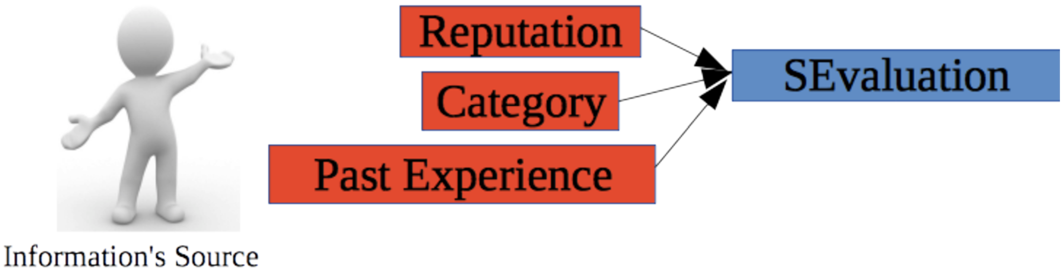

Figure7: Input parameters in order to produce SEvaluation

Certainty and Identity

Computing the general trust on the Source concerning P is a good starting point. However it is not enough. In fact, while this value represents an a priori evaluation of how much a source S is trustworthy, there are other two factors that can influence a trust evaluation.

The first one is the **S's degree of certainty about P (***Certainty***)**. The information sources not only give the information but also their certainty about this information. The same information can be reported with different degree of confidence ("I am sure about it", "I suppose that", "it is possible that" and so on).

Of course we are interested in modeling this certainty, but we have to consider that through the trustor's point of view (it subjectively estimates this parameter). It is defined as a real value in range [0,1].

The second dimension represents **the trustor's degree of trust that P derives from S (***Identity***)**: the trust we have that the information under analysis derives from that specific source; it is defined as a real value in range [0,1]. This parameter has a twofold meaning:

1. For instance, considering the human communication I can be more or less sure that the specific information under analysis has been reported by the source S. It is a problem of memory, do I recall properly?
2. In the web context the communication's dynamics changes. I will probably receive the information by someone hiding beyond a computer. How may I be sure about it's identity? Can I trust that S is really who is saying to be? This is a very complex issue and its solution has not been completely provided by computer scientist.

The source Evaluation is softened by the Certainty and the Identity parameters, since we considered them as two multiplicative parameters. The output of this operation is the actual trust that the trustor has on S:

$$TrustOnS = SEvaluation * Identity * Certainty$$

PDF: the reported information

With the PDF (Probability Distribution Function) we represent the probability distribution that the source reports concerning the belief P.

Given a fixed number of outcomes, which depends on the nature of the information and on the accuracy of the source in reporting the information, with the PDF a source S reports how much it subjectively believes possible each single outcome.

Of course the source can assert that just one of them is possible (100%) or it can divide the probability among them.

Figure 8 shows an example of what we mean with the term PDF. It is divided in slots, each one representing a possible outcome.

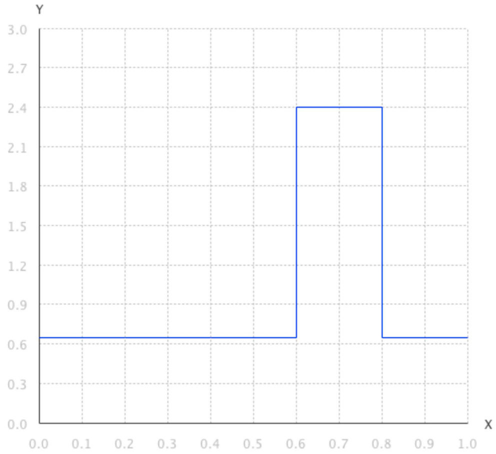

Figure 8: An example of a PDF

It is not possible to consider the PDF as it is. The idea is that if I think I am exploiting a reliable source, than it is good to take into account what it is saying. But if I suppose that the source is unreliable, even if it is not competent or because there is a possibility it wants to deceive me, then I need to be cautious.

Here we propose an algorithm to deal with this problem, combining the trust evaluation with what the source is reporting. In other words, we exploit the $TrustOnS$ value to smooth the PDF. The output of this process is what we call the Smoothed PDF (SPDF).

Recalling that the PDF is divided into segments, this is the formula used for transforming each segments:

$$Segment_i = 1 + (Segment_i - 1) * TrustOnS$$

If $Segment_i > 1$ it will be lowered until 1. On the contrary, if $Segm\square nt_i < 1$it will tend to increase to the value 1.

We will have that:

1. The greater $TrustOnS$ is, the more similar the SPDF will be to the PDF; in particular if $TrustOnS$ =1 => SPDF =PDF;
2. The lesser it is, the more the SPDF will be flatten; in particular if $TrustOnS$ =0 => SPDF is an uniform distribution with value 1.

The idea is that we trust on what S says proportionally to how much we trust S. In words, the more we trust S, the more we tend to take into consideration what it says; the less we trust S, the more we tend to ignore its informative contribution.

Figure 9 resumes the model until this point.

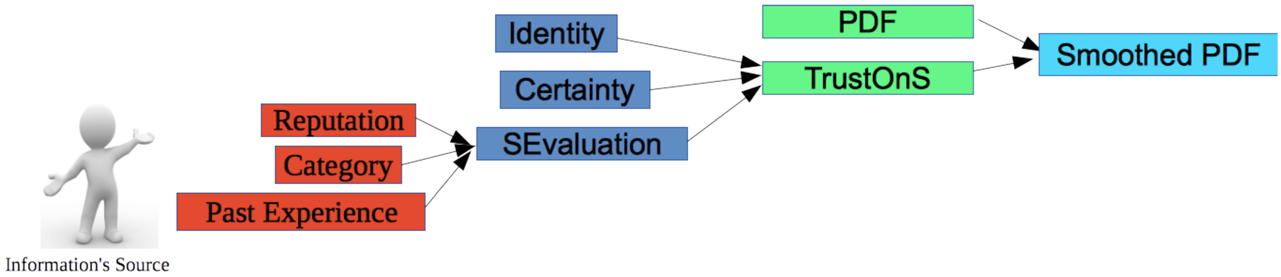

Figure 9: A scheme of the computational model until the SPDF

The effect of each source/evidence on the Global PDF

We define GPDF (Global PDF) the evidence that an agent owns concerning a belief P. At the beginning, if the trustor does not possess any evidence about the belief P, the GPDF is flat, as it is a uniform distribution. Otherwise it has a specific shape the models the specific internal belief of the trustor.

Each information source provides evidence about P, modifying then the GPDF owned by the trustor. Once estimated the SPDFs for each information source, there will be a process of aggregation between the GPDF and the SPDFs. Each source actually represents a new evidence E about a belief P. Then to the purpose of the aggregation process it is possible to use the classical Bayesian logic, recursively on each source:

$$f(P|E) = \frac{f(E|P) * f(P)}{f(E)}$$

where:

f[P|E] = GPDF (the new one)

f[E|P] = SPDF;

f(P) = GPDF (the old one)

In this case f(E) is a normalization factor, given by the formula:

$$f(E) = \int f(E|P) * f(P)\, dP$$

In words the new GPDF, that is the global evidence that an agent has about P, is computed as the product of the old GPDF and the SPDF, that is the new contribute reported by S.

As we need to ensure that GPDF is still a probability distribution function, it is necessary to scale down it[4]. This is ensured by the normalization factor f(E).

Figure 10 represents the whole model for managing trust on information sources.

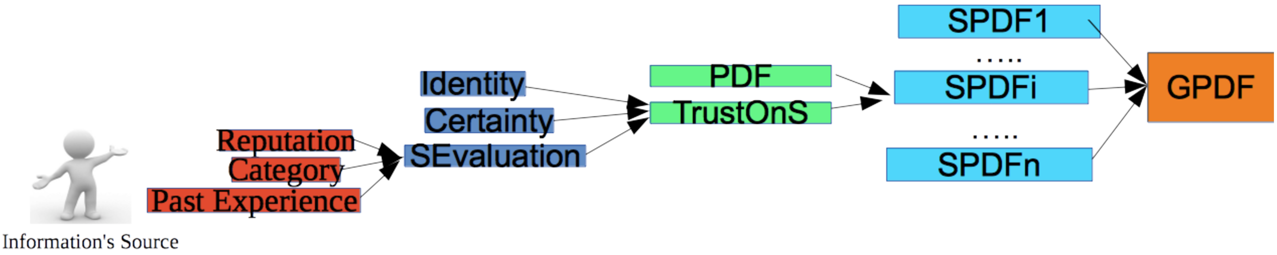

Figure 10: A scheme of the computational model until the GPDF

Exploiting the GPDF, the trust is able to understand what is the outcome O_i that is more likely to happen.

[4] To be a PDF, it is necessary that the area subtended by it is equal to 1.

Handling uncertainty

Dealing with information, a critical point is how to handle uncertainty. The point is that considering uncertainty on information is correct, but it is a too limitative approach. In fact uncertainty comes up at different levels and has to be taken into account when deciding.

Actually, in this model we handle it in three different ways.

The first one is the **uncertainty on the source**. This is given by the source evaluation $SEvaluation$.

The second level is represented by **uncertainty on communication**. This is handled by the two parameters Certainty and Identity: how much I'm sure about the identity of the source? How much certainty does the source express in reporting the information (according to the trustor)?

The last level is the **uncertainty on the reported information** (PDF). This is managed just by the intrinsic nature of the PDF. In fact what happens here is that the source express its certainty/uncertainty through the outcomes' distributions.

In practice, we take into account uncertainty in all the process, until the end, in order to produce a proper prediction.

A Workflow's Example

In this section we want to provide a working example of how to use the model. As the trust computation is quite simple and intuitive, below we will directly use the TrustOnS parameter, together with the corresponding PDF.

Moreover, we will represent PDFs as a list of five values, with the following formalism:

$$PFD_{Si} = [x_{i1}\ x_{i2}\ x_{i3}\ x_{i4}\ x_{i5}]$$

in which x_{i1} x_{i2} x_{i3} x_{i4} x_{i5}[5] are the values of the PDF for the source S_i in the corresponding segment.

Suppose that an agent has to understand what kind of weather there will be the following day. It starts collecting forecast from its information sources. The possible outcomes are five: {sunny day, cloudy day, light rain, heavy rain, critical rain}.

Let's suppose that Source S1 has a TrustOnS_{S1}=1 (the maximal value) and that it is asserting PDF_{S1} = [0.5 0.5 0.5 3 0.5], so it mainly suppose that there will be heavy rain.
The visual representation of PDF_{S1} is provided by figure 11.

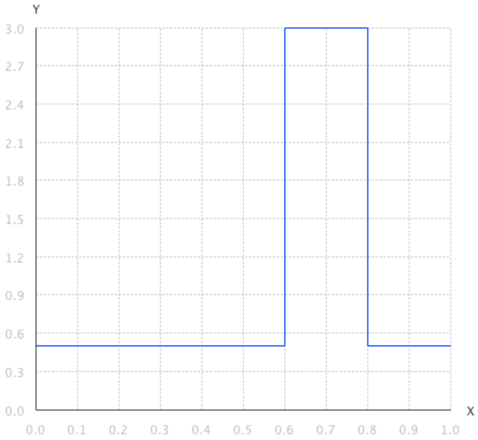

Figure 11: The representation of PDF_{S1} in the example

As the trustor has the maximal trust on S1, PDF_{S1} and SPDF_{S1} will be the same. Plus, as this the first evidence on P, even the GPDF is equal to PDF_{S1}.

[5] Note that, from how the PDF has been defined, these parameters are non-negative real numbers, with the peculiarity that their sum is equal to 5.

Let then see what happens to S2, asserting the same of S1, but with a TrustOnSource$_{S2}$ of 0.7. The PDF$_{S2}$ is the same of PDF$_{S1}$, but the SPDF$_{S2}$ is different, as showed by figure 12:

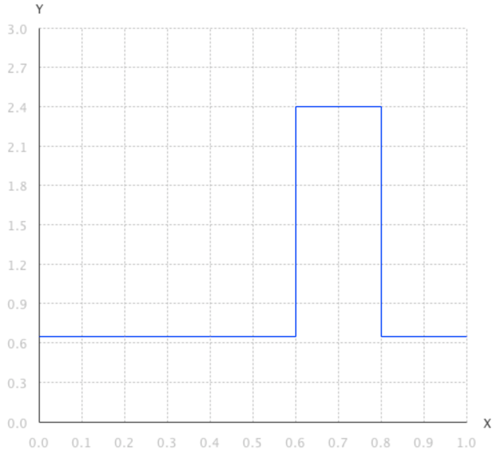

Figure 12: The representation of SPDF$_{S2}$ in the example

The PDF$_{S2}$ has been smoothed, so that values grater than 1 has been decreased and values smaller than one has been increased.

Let's then see what happens to the GPDF:

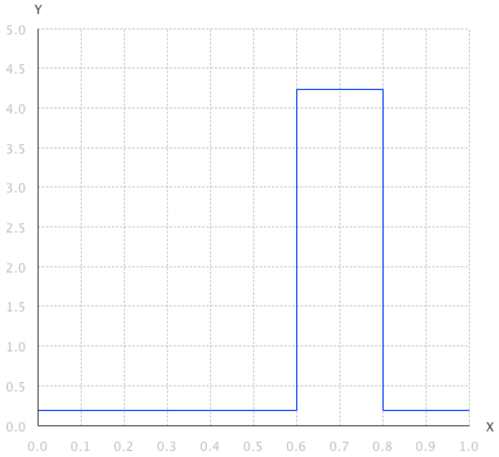

Figure 13: The representation of GPDF of the example, with the contributes of S$_1$ and S$_2$.

As showed by figure 13, Thanks to the fact that the sources, even if with two different trust degrees, are asserting the same things, there is a reinforcement of evidence in segment 4 of the GPDF.
This is a peculiarity that we shaped in our previous models and that persist in this one as a consequence of the Bayes theorem.

Let's than see what happen in presence of a third source S3, with TrustOsSource$_{S3}$ = 0.3 and PDF$_{S3}$ = [0.3 3.8 0.3 0.3 0.3].
This source is reporting a cloudy day forecast. Its SPDF will be:

The final result is showed by figure 14:

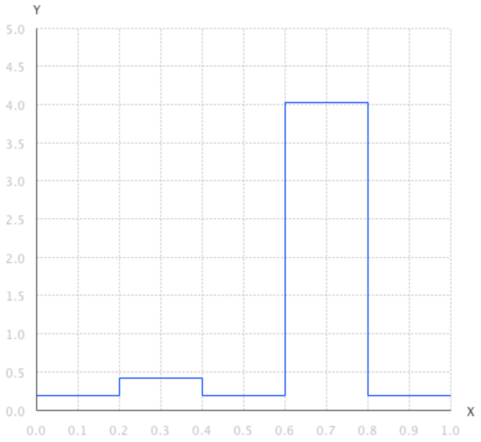

Figure14: The final representation of GPDS in the example

The new GPDF is quite the same of the previous one. This is due to the fact that, although S3 is strongly disagreeing with S1 and S2, it has a low level of trust. Then it will lightly affect what the trustor thinks.
In the end the trustor can assert that there will be heavy rain the next day.

Conclusion

The aim of this work was that of realizing a theoretical and computational model for dealing with information sources.

This is in fact an uneasy task and there can be critical situations in which agents have to face sources asserting different things.

We decided to realize a model as generic as possible. Doing so, the model does not depend on a specific context and it can be applied on different practical context.

The basic idea is that using trust on information sources is a promising way to face the problem. Then, from a theoretical point of view, we analyzed all the possible cognitive variables that can affect trust on an information source.

After analyzing the various ways to represent information, we decided to exploit Bayesian theory. Then we showed how to apply the trust evaluation on the information layers in order to properly take into account information.

Finally, we proposed a practical problem – the one of weather forecast – and we showed how to apply the model in order to get a solution.

CHAPTER 3:

THE ROLE OF CATEGORIES FOR TRUST IN INFORMATION SOURCES.

In this work we are interested in the fact that the relevance and the trustworthiness of information acquired by an agent X from a source F, strictly depends and derives from the X's trust in F with respect the kind of information. In particular, we are interested in analyzing the relevance of F's category as indicator for its trustworthiness with respect to the specific informative goals of X.

In this work we use the cognitive model for trust on information sources proposed in the section "**THE MODEL WITH MULTIPLE SOURCE ASSERTING (OR DENYING) A SINGLE BELIEF**" of chapter 2.
In particular we will focus on the concept of category, as each agent in the world will be considered as belonging to a specific agent's category. We also consider some kind of variability within the canonical categorial behavior and their consequent influence on the trustworthiness of provided information.

The introduced interactive cognitive model also allows evaluating the trustworthiness of a source both on the basis of its category and on the basis of the past direct experience with it, so selecting the more adequate source with respect to the informative goals to achieve. We present the computational approach and some selected simulation scenarios together with the discussion of their more interesting results.

This work has resulted in the following publications:

1. Falcone, R., Sapienza, A., & Castelfranchi, C. (2015). The relevance of categories for trusting information sources. *ACM Transactions on Internet Technology (TOIT)*, *15*(4), 13.
2. Falcone, R., Sapienza, A., & Castelfranchi, C. (2015). Trusting Information Sources Through Their Categories. In Advances in Practical Applications of Agents, Multi-Agent Systems, and Sustainability: The PAAMS Collection (pp. 80-92). Springer International Publishing.
3. Falcone, Rino, Alessandro Sapienza, and Cristiano Castelfranchi. "Valorizing Prejudice in MAS: A Computational Model", proceedings of the workshop COOS 2015, In Conjunction with AAMAS 2015.
4. Falcone, R., Sapienza, A., & Castelfranchi, C. (2016). The predictive role of prejudice: a computational model for using categories. *International Journal of Computational Intelligence Studies*, *5*(2), 121-137.
5. Falcone Rino, Alessandro Sapienza, and Cristiano Castelfranchi. "Trusting through categories", in proceedings of the conference AISC mid-term 2015 Palermo, Nea Science year 2, volume 7, 53-55

1. Relevant questions about the categorization of the agents and their trustworthiness

What is a category? We can define a category as a set of features defining their functional/symbolic/descriptive structure. These definitional features should be present in each of the elements belonging to that category. In this work we assume that the membership to a given category is assigned in a top-down and transparent way; then, the category of a given agent is public, common knowledge.

The advantage of such a *hierarchic structure of knowledge* is not only economical: in memory space and costs; the greatest advantage is in the fact that we *know a lot of things about something/someone that we never met*; just by inference, prediction, inheritance.

We have a lot of knowledge about a given entity without any direct experience on it. This crucial power of our cognitive organization is obviously exploited also in social life, in order to have information and expectations about people that we never met.

Some very relevant aspects about trusting categories (and agents belonging to them) can be resumed as follow:

- **Categories for competence and categories for reliability**: we will use both these kinds of categories, and in particular we have chosen *taxi-driver*, *policeman*, *passer-by* and *shopkeeper* as categories for the competence dimension (agents belonging to these categories have common features of competence deriving from the same definition of these categories: for example, a policeman has competence about security questions and public order, a taxi-driver about city directions and traffic, and so on). At the same time we have also chosen three abstract categories relative to the reliability: reliability due to the *role* of the agent (in some sense all the agents belonging to a specific category should have that feature of reliability: for example a policeman -for role- has to be sincere and motivated about security information); reliability due to the *individuality* of the agent (a reliability feature due to the specific agent, independent from its category: this is a trait of personality; it is implemented through an expressivity mark of the agent); reliability due to the *culture of the* *environment* in which the agent is absorbed (a reliability feature due to the specific cultural environment that affects all the categories and the agents in that environment).

- **Mixed categories**: In fact, as just considered, we have to put together the different features of several categories. Not only because we have categories for both competence and reliability, but also because we should mix different categories of competence and of reliability: an agent might belong (and generally belong) to more than one category. In the second simulation we inquire these aspects.
- **Homogeneity and stability of the trustworthy features of the agents belonging to the same categories**: Are the agents belonging to the same category (for example all the policemen) equally trustworthy with respect to a specific information request? How big is their variability? To what extent this variability determines a stable or unstable trustworthiness for that category with respect that informative task? See the fourth simulation for a specific analysis of this aspect.

2. Computational model

In order to realize our simulations, we exploited the software NetLogo (Wilensky, 1999). It is an open source agent-based programming environment, particularly suited for modeling natural and social phenomena.

General Setup

The simulations were carried out using (except one case) 40 trustees and 1 trustor.

In particular, we decided to classify trustees into 4 categories: shopkeepers (Sk), passers-by (Pb), taxi-drivers (Td) and policemen (Pm).

As usual in the use of categories, *agents belonging to them inherit with a certain regularity the features attributed to those categories*. In our case given a specific type of required information, the agents belonging to each category can perform differently, in terms of competence and reliability. The variability of the behavior within each category is ruled by the uncertainty factors (both for competence and reliability). In other words, the agent's performance on a specific task mainly depends from the agent's category (both its category of competence and its category of reliability) and secondarily from the specific features of that agent (that express its variability within the category).

As we will see later it is also important to model not only the top-down link (inheritance) but also the reverse process: how the evaluated trustworthiness of a given member of a category bottom-up builds or affects the trustor's opinion on its category. The trustors have evaluation on categories by reputation, recommendation, or analogy, or higher categories; but they also build or adjust it on the basis of their direct experience; however direct experience is with individuals, as members of that category.

Then each category will be characterized by:

- **competence**, in range [0,100];
- **uncertainty on competence**: we fixed this value on 20%;
- **reliability**, in range [0,100];
- **uncertainty** on reliability: we fixed this value on 20%;

We fixed top-down these values. In fact, we are not interested in this work in showing how these values should be computed (even if in fifth simulation we show a way to do that), but we assume that them are common knowledge, accessible to the trustor.

As a consequence of these parameters, each trustee will be characterized by;

- **competence**, in range [0,100], derived from the value of the belonging category;
- **reliability**, in range [0,100], derived from the value of the belonging category, but also influenced by **individual_reliability** and **contextual/cultural atmosphere** (see below);
- **individual_reliability**: this is a peculiar feature of each trustee; it could be more available than expected, given its category (+20% on reliability), neutral (no influence on reliability) or less available (-20% on reliability). This trait is expressed by a visual feature that the trustor can access: it is a perceivable feature of the trustee. We modeled it in NetLogo changing the image of the trustees: happy face, neutral face and sad face;
- **contextual/cultural atmosphere** of trustees: it is an additional parameter that influences the reliability of all the trustees in the same way;
- **own trustworthiness** (objtw), in range [0,100], given by the mean of competence and reliability; it represents the objective probability that, concerning a specific kind of required information, the trustee will communicate the right information;
- **past experience**: this value represents how the trustee performed with the trustor concerning a specific kind of required information; this value is obtained through the experience vector;
- **experience vector**: a vector in which the last performances of the trustee are stored; they are boolean values (as in our cases the information can just be true or false, the trustee can just confirm or deny it).

Kind of Information

First of all, it is important to underline that in each simulation the information can be just true or false (and that the correct information is always the true one). Given this assumption, we defined 4 different types or categories of required information (they represent the different tasks to achieve by the trustees):

- **ask for x**: to this kind of request just one category of trustee will perform badly, all the others will perform properly;
- **ask for y**: to this kind of request just one category of trustee will perform properly, all the others will perform badly;
- **ask for z**: to this kind of request all the trustees will perform properly (an example could be "ask for hospital"); in cases such this the trustees are conditioned by cultural/moral factors on reliability;
- **ask for t:** this is a particular kind of request, in which trustees perform all in the same way regardless of the category to which they belong: 100% of trustworthiness and zero uncertainty;
- **ask for j:** to this kind of requests, the performances of the different categories present a variability of response with respect to each informative task less evident that in the previous cases (x and y) and, as a consequence, determine the various behaviors of the belonging trustees (examples could be "ask for street out of the neighborhood", "ask for parking area", "ask for shops' opening hours";);
- **ask for k:** this is a kind of information request in which is unpredictable the performance of the agents' categories, so that the trustor doesn't know how the trustees' categories will perform: it ignores the trustworthiness of all the categories about that specific request of information.

Weights of Category and Past Experience

According to the simplification of our model, we consider just “past experience” and “category/analogy” as dimensions for evaluating a trustee as information source (we exclude reputation/recommendation).

It is important to assign weights to these dimensions, to establish which of them is more relevant in different situations and scenarios.

We chose to compute them as follow:

- **the past experience weight** depends just on time: it is increased of 1 unit every tick of the simulation, starting from 0 and arriving to a maximum value of 10. This models the fact that, the more the trustor experiences the world, the more this experience acquires importance. In any case if the experience is too old (more than 10 steps), then it will not have any role.
- **the category weight** depends on the mean value of uncertainty on the dimensions of each category (to whom the agent belongs): as for each dimension of each category we give an uncertainty of 20%, this weight is fixed to 8 (10 – 2).

How the Simulations Work

In general, what happens is that the trustor moves around the world and, in every tick, meets a number of trustees. It asks about P (the information it needs) just to its neighbors (trustees with distance less than 3 NetLogo patches) and it also evaluates them. For the evaluation we use two different approaches, comparing their performances:

- the first one uses just past experience;
- the second one exploits both past experience and category.

Detailed settings

In simulations world there are **4 trustees categories**. Each one, on the base of the kind of required information, is characterized by the two parameters **competence and reliability**, both linked to an **uncertainty value** (fixed to 20%, when it is not specified). The trustees' membership to categories is clear and non-misleading; the trustor does not determine it, but it is a ready-to-use information.

The categories' settings (just those we used) for each information request are:

1. **Ask for y (simulation 2):** Sk: competence= 10, reliability = 80; Pb: competence= 10, reliability = 50; Td: competence= 10, reliability = 80; Pm: competence= 90, reliability = 70; contextual/cultural atmosphere = 0;
2. **Ask for t (simulation 3):** Sk: competence= 100, reliability = 100; Pb: competence= 100, reliability = 100; Td: competence= 100, reliability = 100; Pm: competence= 100, reliability = 100; contextual/cultural atmosphere = 0; Also the uncertainty for all the categories is 0%.
3. **Ask for j1 (ask for street out of the neighborhood, simulation 1):** Sk: competence= 30, reliability = 30; Pb: competence= 40, reliability = 50; Td: competence= 90, reliability = 20; Pm: competence= 70, reliability = 90; contextual/cultural atmosphere = 0;
4. **Ask for j2 (ask for parking area, simulation3):** Sk: competence= 50, reliability = 50; Pb: competence= 50, reliability = 50; Td: competence= 50, reliability = 50; Pm: competence= 90, reliability = 90; contextual/cultural atmosphere = 0;
5. **Ask for j3 (ask for shop opening hour, simulation 4):** Sk: competence= 90, reliability = 90; Pb: competence= 50, reliability = 50;

Td: competence= 30, reliability = 30; Pm: competence= 50, reliability = 70; contextual/cultural atmosphere = 0;

6. **Ask for k (simulation 5):** Sk: competence= 10, reliability = 10; Pb: competence= 30, reliability = 30; Td: competence= 50, reliability = 50; Pm: competence= 70, reliability = 70; contextual/cultural atmosphere = 10.

Starting from these values, we generate competence and reliability for the trustees.

These values are generated through a gaussian distribution, using as mean the corresponding category value and as variance the uncertainty associated to it.

The reliability value depends also on other two parameters:

1. **individual_reliability:** it randomly affects the trustee's reliability in a positive (+20%), negative (-20%) or neutral (+0%) way.
2. **contextual/cultural atmosphere** of trustees: it is an additional parameter that influences the reliability of all the trustees in the same way; it depends on the kind of the information request.

Then the reliability of a trustee is given by:

final_reliability = reliability + individual_reliability + contextual/cultural_atmosphere

Finally, the most important parameter is the **own trustworthiness**, that determines the behavior of a trustee. It is obtained by the mean of competence and reliability (actually the final_reliability) of the trustees.

Concerning the trustor, it is characterized by the following variables:

1. **length of experience vector**: from 1 to 200 in the first experiment, but generally set to 10;
2. **the past experience weight:** starting from 0, it is increased of 1 unit every tick of the simulation, arriving to a maximum value of 10;
3. **the category weight:** depends on the mean value of uncertainty on the dimensions of each category (to whom the agent belongs): as for each dimension of each category we give an uncertainty of 20%, this weight is fixed to 8 (10 – 2);
4. **an evaluation for each trustee met:** having the information on past experience and category as fuzzy value, the trustor estimates an evaluation exploiting the centroid method (previously described). It can decide to exploit these dimension both, or just he past experience.

The main workflow of the simulations can be summarized as:

1. A new world (17 x 17 patches) is generated, together with agents belonging to it (trustor and trustees). The number of agents changes on each simulation;
2. the trustees and the trustor move around the world of one patch in a random direction;
3. then the trustor ask to its neighbors about P. They will assert 1 (correct answer) with a probability that corresponds to their own trustworthiness; otherwise they will assert 0 (wrong answer);
4. the trustor memorize each answer, updating the evaluation of its neighbors.

Outputs

In every simulation we use some different indexes to understand the result of the simulation. For example:

1. **Mean Error of Evaluation (MEV)**: for each tick of the simulation, we compute the mean absolute error of evaluation of the neighbors. The mean error of evaluation is a global result that represents the mean of all mean absolute errors of evaluation of the neighbors in time. This is computed with both the evaluation's algorithms described above; here is a formula to better explain this:

$$\frac{\sum_{j=1}^{T} \sum_{i=1}^{N_j} |objtw_i - eval_i|}{\sum_{j=1}^{T} N_j}$$

where: $objtw_i$ is the objective trustworthiness of i-th trustee; $eval_i$ is the evaluation, according to the trustor of i-th trustee, obtained exploiting the two dimension of category and past experience (what we previously defined as $MainTrust_X(F,P)$); T is the number of current ticks; N_j is the number of neighbors at tick j.

MEV (based on absolute error) is particularly useful to provide an estimation of how much the evaluation produced by the trustor differs from the effective objective trustworthiness of the trustee on average.

It is worth underlining that we are not interested in showing the distribution of evaluations, as it is linked to the category mean trustworthiness, that changes from trustee to trustee. We prefer to use an error index able to show the efficacy of evaluation's algorithms. Precisely for this reason, we chose to use MEV.

2. **Mean error of evaluation for a given category**: it is the same of MEV, but here we consider just the contribute of a single category.
3. **Success rate**: it is the percentage of success of a category given a kind of information request.

3. SIMULATIONS

First Simulation: the relevance of the memory's length

In this simulation we investigate the size of experience vector (VS), in words, how much past experience the trustor has to consider in order to evaluate better the trustees. Normally it is fixed to 10, but what does it happen if it changes? We tried to reduce it and to increase it, for understanding the effect of these parameters.

Of course, we expect that this variation on the considered window of storage will influence somehow the value of the past experience.
How does this change influence the MEV? Can we identify the limits beyond which we are storing too little information or too many information?

For this simulation we set:

1. kind of information: "ask for j1";
2. number of trustees: 10 for category;
3. number of ticks: 400;

We want to find out the *minimal window* for the past experience vector beyond which the error increases too much and the *maximal window* after which the error doesn't decrease.

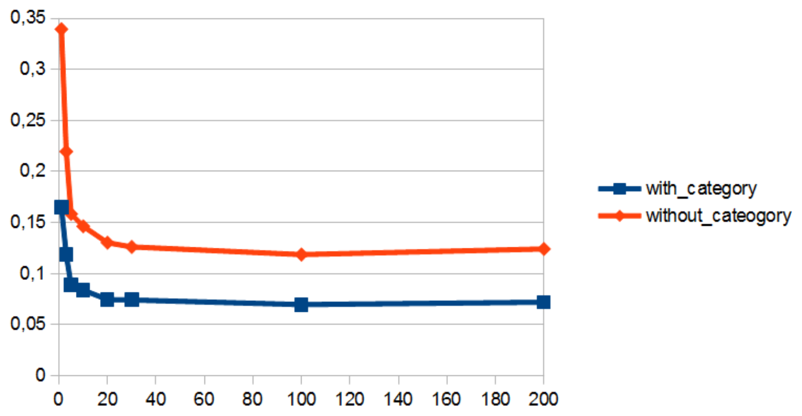

Figure 15: Representation of MEV and MEV without category

Figure 15 shows a graph that exposes the result of the simulations. On the abscissa we have the VS and on the ordinate the MEV, in particular the two curves represent the MEV (in blue) and the MEV without category (in red). We can clearly see how the two errors decrease when the vector size increases.

With VS=20, we have a good improvement, especially for MEV without category. Instead, the improvement from 20 to 30 is quite nothing for the evaluation's algorithm that uses categories and less than 0.5 for that based just on past experience.

Following we have that, although we increase the VS, there is no improvement from a length of 100 to a length of 200: actually there is a worsening (due to statistical randomness into the model). It means that we reach the minimal error close to 100.

Likewise, let's see what happens when it is reduced the VS (moving up the curve towards zero): there is a gradual increase of the error of evaluation that reaches its maximal value when we set the VS to 1, the lowest value.

In the previous example the trustees' trustworthiness doesn't change in time. In these cases it is obvious that the more information on past experience we have, the more precise will be the evaluation about their trustworthiness.

Let’s now investigate the case in which the trustworthiness changes, considering three different VSs: 3, 10 and 100. Suppose trustworthiness changes every 30, 100 and 300 ticks (we called it PTC, i.e. **period of the trustworthiness’ change**).

For this simulation we set:

1. kind of information: “ask for j1”;
2. number of trustees: 10 for category;
3. number of ticks: 3000 (it is necessary more time for evaluating the effects of trustworthiness's dynamics);

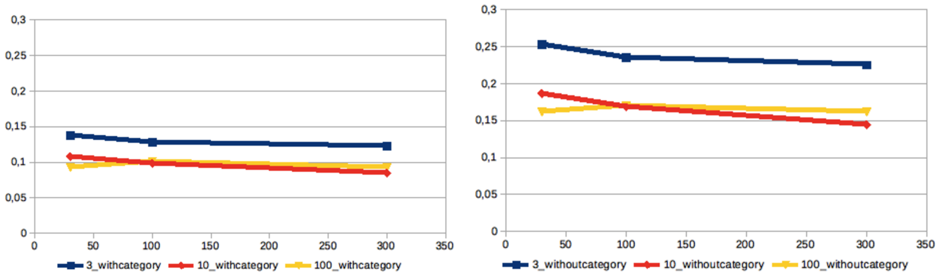

Figure 16: Representation of MEV(on the left) and MEV without category (on the right). The abscissa represents PTC, while the ordinate represents the MEV for vector sizes 3, 10 and 100

It is clear from Figure 16 that the MEV increases when the period decreases: the longer is the VS, the greater is this decrement.

Let’s then consider the case in which the trustworthiness' dynamics gets as possible values just 100 and 0 (limit case) and it changes every 30 ticks.

Table 11. MEV and MEV without category when the trustworthiness changes every 30 ticks

	VS = 3 **MEV with category**	VS = 3 **MEV without category**	VS = 10 **MEV with category**	VS = 10 **MEV without category**	VS = 100 **MEV with category**	VS = 100 **MEV without category**
Average	0,1853	0,2992	0,2426	0,3974	0,2775	0,4524

As easily predictable, given the chosen dynamics of the trustworthiness, the shorter is the VS, the better is the MEV. Also interesting is the analysis of the results: for example how MEV changes by increasing the VS (and differently with or without category).

In conclusion, we can say that we have:

- **a too short memory**: when the experience vector fails to shape properly the past experience of the trustee; in practice the cumulated experience of the trustee's behavior is not enough for well representing it.
- **a too long memory**: when it memorizes too many old experiences; it could include not more current behaviors of the trustee: this is the case in which the trustee's trustworthiness changes in time.
- **a right memory**: when it memorizes just the needed quantity of information: neither too much nor little; the information included in the experience vector is enough for both representing all of the trustee's behaviors and, at the same time, for taking account for changes of its trustworthiness.

These results, given the chosen model, were quite predictable. The added value is in the confirmation of the hypothesis and in the definition of the parameters. For example, in the experiment seems that a good compromise is represented by VS = 10.

Second Simulation: trustworthiness of mixed categories

In this setting we experiment the behavior of mixed category of trustees.

We consider the case in which the trustor perceives all the agents it meets only as "passers-by", while in the reality:

- 10 are just passers-by;
- 10 are both passers-by and police-men;
- 10 are passers-by and shopkeepers;
- 10 are passers-by and taxi drivers.

What does it mean in our terms to belong to a mixed category? It means that we have to build the value of its resulting trustworthiness starting from the starting ones. In particular we consider that the trustee's competence is the highest among the categories to which the trustee belongs (in fact it has both the competences, so it is natural to use the best one); while the reliance is the one in which the trustee is playing the role. Here we have to clarify that with the reliance dimension we intend all those features related with the trustee's motivation (reliability, intention, willingness, and so on). Some of these features are strictly regulated/obliged by the role and not necessarily spontaneously practiced by the agent out of the role. In this sense, on the contrary of the competence dimension, we consider the trustee's reliance related to the attributed role to it.

Let's suppose that trustee Y is a passer-by, but also a policeman. When the trustor asks for information to it (as a passer-by), its reliability will be the same of passers-by, because in that moment it is a passer-by and it will act accordingly (it could also operate with the reliance of a policeman, but this cannot be predicted and practiced nor necessarily known by the trustor).

But it is not possible to state the same for its competence. If Y (for the kind of required information) has a higher competence as policeman that as passer-by, its competence will not decrease because of its current role. In fact for mixed categories we consider the competence always as the maximal between the originating ones (it seems a reasonable hypothesis).

Then we'll have that trustees can perform better than expected by the trustor that perceives them as belonging to only one category: in fact, its information about the trustees' categories is incomplete.

In this scenario we investigate two points:

- *What is the difference between a situation in which the trustor perfectly knows the categories of the trustees and the one in which it only perceives partially their categorization (case of the mixed and hidden categories)?*
- *Accordingly to the trustees' behavior, can the trustor cluster them? How can the trustor rebuild the correlation between these new clusters and the mixed categories?*

The setting for these simulations is:

1. kind of information: "ask for y"
2. number of trustees: 10 for category (mixed categories);
3. number of ticks: 400;

Let's see the results:

Table 12. Comparison between a simulation with mixed categories and one with normal categories

	MEV with category (mixed categories)	**MEV without category (mixed categories)**	**MEV with category (normal categories)**	**MEV without category (normal categories)**
Average	0,1039	0,1451	0,078	0,1368

We can clearly notice that the MEV is greater in the case of mixed categories, because the prediction about the behavior of these categories is less precise. The "MEV without category" is quite the same in both cases, because here it is taken into account just the past experience, which shapes the real behavior of trustees.

Although the trustor sees the trustees as belonging to the same category, by the means of a clustering process it is possible to classify them on the basis of their behaviors: trustees will be categorized according to their performance (excellent, good etc.). We can clearly notice a matching between the original categories and these new ones: in fact policemen emerge from the other trustees, having a higher competence on the chosen task.

Table 13. Results of the clustering process

	excellent	good	mediocre	poor	terrible

Passers-by and shopkeepers	0	3,3%	23,7%	31%	0
Passers-by	0	6%	20,5%	31,5%	0
Passers-by and policemen	0	84,9%	34,7%	3,1%	0
Passers-by and taxi-drivers	0	5,8%	21,3%	33,9%	0

Let us develop a brief consideration. The failing results shown in Table 12 could induce the trustor to verify the real category of the trustee. In this sense the opportunity of applying the clustering process presented in Table 13 indicates the potential solution to its misclassification.

We have also to underline that the definition of different categories in which classify the agents, both on the basis of the same or different dimensional traits, offers the possibility of both better understand and predict the agent’s behaviors.

Third Simulation: convergence speed of the evaluation’s algorithms

In this simulation we want to understand which of the two evaluation's algorithms (the one with category and past experience and the one just with past experience) performs better; in other words, which of them can provide a better evaluation in the shortest time.

Here the trustor evaluates the performance of the trustees it meets. For each trustee, it memorizes the last three evaluations (we are considering the evaluation, not just the performance of the trustee, which only describes success or failure) both for the two evaluation's algorithms.

To this purpose, it is necessary to define what we intend about convergence. We can say that there is convergence with an algorithm on a given trustee when the last three evaluations computed with this algorithm are equal and correct (namely the fuzzy value of the evaluation corresponds to the fuzzy value of the trustee's trustworthiness).

When there is convergence with one algorithm on a trustee, this trustee is marked, so that it won't be considered for the rest of the run and we give one merit point (mp) to the converging algorithm. The simulation is considered ended when all the trustees has been marked. Presented results are the average of 10 runs.

The setting for these simulations is:

1. kind of information: “ask for j2”;
2. number of trustees: 10 for category;
3. number of ticks: as many as needed for all trustees to converge

Table 14. Performance of the two evaluation's algorithms

	Converged without category	Converged with category	Equally converged	Needed ticks
Average	1,7 mp 4,25%	15,8 mp 39,5%	22,5 mp 56,25%	309,4

"Needed ticks" represents the timed needed for the simulation to end.

As expected, we have that the algorithm that exploits both category and past experience performs better than the other one.

However there is a special case. Suppose that the categories are equally adequate with respect to the informative task (all good, or mediocre, or bad, and so on). What would happen?

The setting for these simulations is:

1. kind of information: "ask for t"
2. number of trustees: 10 for category;
3. number of ticks: as many as needed for all trustees to converge

Table 15. Performance of the two evaluation's algorithms

	Converged without category	Converged with category	Equally converged	Needed ticks
Avera	0 mp	1 mp	39 mp	367,7

ge		(2,5%)	(97,5%)	

In this situation the two algorithm perform quite the same (actually there is a difference in the needed ticks, but it is too high and it is due to the random nature of the simulations), meaning that there is no reason to consider the categorical nature of the trustees. In our case this additional information on the trustees is free of cost but in general, one should take into account that accessing to this value could have a cost.

In sum, exploiting categories is quite always convenient in term of time, as one can obtain a good trust evaluation in less time, but there could be situation in which, considering also the cost associated with this information, it could be better to rely just on past experience.

Fourth Simulation: corrupted categories

What if some of the trustees are not representative in the performance with respect to their category? Is the trustor still able to identify good trustees? Does this framework still work?

In this simulation we introduce a given percentage of trustees performing better or worse than they should do given their membership to some categories.

In particular, we chose a percentage of 80%[6] of false trustees (non representative of the category), which will have their own trustworthiness,

[6] Actually, in the case of positively corrupted categories, the percentage is 60%: shopkeepers will not be affected by this change, as their trustworthiness can be 90% for this kind of information. This is the reason why the trustor chooses the 84% of time an old trustee: the 60% of trustees is positively corrupted, while the 25% has a high value of trustworthiness even if it isn't corrupted.

increased to 90% in a simulation or decreased to 10% in another simulation.

In this simulation we completely change the experimental approach.

In a first step the trustor explores the world, making experience with the trustees it meets. It will find good trustees that perform according to their category, and bad trustees, that perform differently from their category.

Then, in a second step, we introduce other new trustees, with which the trustor had no experience before.

Who will the trustor ask for the information? It will ask for information to new trustees or to old ones?

As reasonable, the result is that the trustor will mostly chose new trustees in the case of negatively corrupted trustees (in which the categories seems not reliable for predicting the behavior of their components), and old trustees in the case of positively corrupted trustees (in which the categories seems more reliable than expected).

The setting for these simulations is:

1. kind of information: “ask for j3”;
2. number of trustees: 5 for category in the first step, with the addition of 5 others in the second step;
3. number of ticks: 200 for the first step, 200 for the second step.
4. Percentage of false negative trustees: 80%

For this experiment we made 10 runs. For each run, after the first step, we verify the behavior of the trustor for 10 ticks, checking if it uses an old or a new trustee as source of information. In this first experiment (negatively corrupted trustees) we expect that the trustor will choose new trustees, so if it doesn’t happen we will explain why.

Table 16. Chosen trustees in presence of negatively corrupted trustees

	Good choice	Bad Choice
Chooses new	66%	
Chooses old	19%	15%

Let's see what happens if we introduce trustees performing better than their category (positively corrupted trustees).

The setting for these simulations is:

1. kind of information: "ask for j3"
2. number of trustees: 5 for category in the first step, with the addition of 5 other in the second step;
3. number of ticks: 200 for the first step, 200 for the second step;
4. Percentage of false positive trustees: 80% (actually it is 60%, as corrupted shopkeepers' trustworthiness still belong to the range of shopkeepers' trustworthiness for this kind of information).

In this second experiment we expect that the trustor will choose old trustees, so if it doesn't happen we will explain why.

Table 17. Chosen trustees in presence of positively corrupted trustees

	Good choice	Bad Choice
Chooses new	16%	
Chooses old	84%	

Let's then analyze this result. For the negatively corrupted categories the trustor chooses:

1. 66% of time a new trustee, as expected: this is the best rational choice as the 80% of old trustee has performed badly;
2. 19% of time an old trustee, but the choice is right (maybe because it chooses a not corrupted trustee or there are just old trustees);
3. 15% of time an old trustee, but the choice is wrong.

In the case of positively corrupted categories the trustor chooses:

1. 84% of time an old trustee, as expected: again, this is the best rational choice, as the 85% of old trustee has performed well[4];
2. 16% of time a new trustee, but the choice is right (maybe because there are just new trustee).

As we can see, in the second experiment the case in which the trustor chooses wrongly a new trustee does not show up (even if it is still possible). Of course this fact is due to the high presence of over-performing old trustees. Also it is unlike to choose someone unexperienced, if the experienced ones performed very well.

Further details and explanation of trustor's choices are listed below.

Table 18. Explanation of trustor's choices

Kind	**New**	**Old**	**Why**
negatively corrupted trustees	6,6	3,4	- Trustor chooses an old trustee because it is surrounded only by old trustees - Trustor chooses an old false trustee that it has never experienced

			- Trustor chooses an old experienced trustee as it is a good member of its category - Trustor chooses an old shopkeeper performing badly and seems reliable rather than a new trustee that has a low category value (taxi driver) and is unreliable - Trustor (wrongly) chooses an old experienced trustee (a shopkeeper) because, even if the past experience is "poor", the evaluation of the category is "excellent"
positively corrupted trustees	1,6	8,4	- Trustor chooses a not experienced trustee given its high category value - There is just a new trustee nearby - The trustor chooses a new trustee because the old trustee, even if it is a good element of its category, had a low performance

Fifth Simulation: trustees without a link with categories

When the value of the category is associated with an informative task, the trustor has a good advantage in the evaluation of trustees belonging to that category, as we showed in the previous simulations.

But there could be situations in which it is not possible to access to this value. In practice, could be possible that the request of specific information is not directly linked with the categories.

In this situation, is the trustor able to establish how each category will perform on this new informative task? In order to obtain this information the trustor has to test categories' performances, in practice the performances of agents belonging to them.

In this simulation we investigate the case in which the trustor is going to ask to trustees a new kind of information, without knowing how the categories perform on this informative task. And so it is useless to know (or to attribute) the categories of membership of those trustees. In this case it can exploit just the direct experience with the trustees.

In order to deduce the evaluation of the category from the past experience, the trustor controls the performance of each trustee, computing the success rate (above defined) of each category. So in this case we have that the trustworthiness of the categories about these new informative tasks would emerge from the bottom of the interactions between trustor and trustees. The evaluation of the category it is just the fuzzy value of the category success rate.

The setting for these simulations is:

1. kind of information: "ask for k"
2. number of trustees: 10 for category;
3. number of ticks: 400.

We made ten runs, reporting in the table 19 the results.

Table 19. Value assigned to categories on the basis of their success rate

	Sk success rate	Pb success rate	Td success rate	Pm success rate

Total	5 x terrible 5 x poor	6 x poor 4 x mediocre	10 x mediocre	10 x good

Trustees were classified:

1. shopkeepers: 5 times as terrible and 5 times as poor; actually they were “terrible”;
2. passers by: 6 times as poor, 4 times as mediocre; actually they were “poor”
3. taxi drivers: 10 times as mediocre, as they were;
4. police men: 10 times as good, as they were;

The percentage of correct classification is 77.5%, and the remaining 22.5% of classification is not to far from the right classification.

This can be considered as a successful result.

4. CONCLUSIONS

In this work we have resumed and better specified the socio-cognitive model of trust in information sources, already analyzed in other previous works (Castelfranchi et al, 2003). We have underlined how this type of trust (in information sources) is in fact just a kind of social trust, ruled by the same theoretical framework including both the dimensions of competence and reliability of the trustee (information source), and the relevant and specific role of the (informative) task.

Also, the reasons to trust follow the same nature of the more general phenomenon of trust: direct experience, recommendations/reputation, logical inference (analogy, categorization, and so on). All these aspects have to be modeled on the information sources’ trustworthiness and on the special relationship between informer (trustee) and informed (trustor).

However there are also sophisticated aspects uniquely related to the trust in information sources: the trust about the self-trust of the sources, the trust that the received information is correct and complete; the trust that the source is really the believed source (problem of identification), and so on.

For reasons of simplification we focused our analysis just on a reduced version of our model of trust in information sources, with a specific attention to the categorization aspects. Our purpose is to show how the categorical aspect is particularly useful for applying the concept of trust in information sources. In fact, categories could both be designed in a top-down approach and emerge in a bottom-up approach through association of similar structural and functional features. So they represent a relevant guide for defining the trustworthiness of the different sources under analysis with respect to the more or less specific informative task.

Many of the results of our simulations were rather predictable. The exploitation of an additional feature of the information sources (their categorical specificity) influencing the behaviors of their members (trustees) determines an obvious advantage for the trustors: They have in fact, additional knowledge about the trustees, also towards never met before trustees. This is especially crucial in an “open world”.

In any case it has been relevant both to confirm with our experiments this advantage and to identify the role of the several parameters involved in the simulations models. For example, how much memorized experience is necessary for choosing a reliable trustee? What variability in the categories is admissible for their fruitful use? How the reference to mixed categories can be useful for trusting their members?

In fact, inquiring these factors, the analysis provided in our simulations offers the possibility of better understanding and predicting the agent's behaviors.

CHAPTER 4:

RECOMMENDING CATEGORIES

With the continuous growth of Internet and the online social networks, the world as we know is in fact developing a dual face. On one side we have what could be defined as “**human society**”; we are used to live there, we know its rules and how to act according to that. But on the other side we have the so called “web society”, a virtual environment that nowadays is becoming a new digital world, affecting our lives and our way to behave.

In this work we focus on the importance of generalized knowledge (agents' categories) in order to understand how much it is crucial in these two worlds. The cognitive advantage of generalized knowledge can be synthesized in this claim: "It allows us to know a lot about something/somebody we do not directly know". At a social level this means that I can know a lot of things on people that I never met; it is a social "prejudice" with its good side and fundamental contribution to social exchange.

In this study we will analyze and present the differences between these two worlds. On this basis, we will experimentally inquire the role played by categories' reputation with respect to the reputation and opinion on single agents: when it is better to rely on the first ones and when are more reliable the second ones.

We will consider these simulations for both the two kind of world, investigating how the parameters defining the specific environment (number of agents, their interactions, transfer of reputation, and so on) determine the use of categories' reputation and trying to understand how the role played by categories will be important in the new digital world.

In this work we use the cognitive model for trust on information sources proposed in the section "**THE MODEL WITH MULTIPLE SOURCE ASSERTING (OR DENYING) A SINGLE BELIEF**" of chapter 2.

This work has resulted in the following publications:

1. Falcone, R., Sapienza, A., Castelfranchi, C. (2015, July). Recommendation of categories in an agents world: The role of (not) local communicative environments. In Privacy, Security and Trust (PST), 2015 13th Annual Conference on (pp. 7-13). IEEE.
2. Falcone Rino, Alessandro Sapienza, and Cristiano Castelfranchi. "The Utility of Categories through their Recommendation in an Agents World with Local or not, in Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science, Ceur workshop proceedings, vol 1419, paper 7.
3. Alessandro Sapienza, Rino Falcone, Cristiano Castelfranchi. "The Positive Power of Prejudice: A Computational Model for MAS, in proceedings of the workshop WOA 2015, Ceur workshop proceedings, vol 1382, paper 6, 39-45.
4. Falcone Rino, Alessandro Sapienza, and Cristiano Castelfranchi. "Exploring Categories Recommendations within human and digital societies", in proceedings of the conference AISC 2015 Genova, Nea Science year 2, volume 9, pp. 75-80. **This work is currently candidate as the best paper of the conference and an extended**

version of it will be published with special mention in the journal "Sistemi Intelligenti".

1. Introduction

Knowing without knowing

Knowledge generalization and its organization around "classes" of entities and events (Falcone and Castelfranchi, 2008) is a foundational need of human cognition. It is not enough to have a scattered knowledge about single dogs or bone, a single act of eating; we need to have non-episodic knowledge about "dogs", "bones", "eating". That's why we categorize objects and facts, and create "classes" and "types" and belonging relation. The advantage of such a *hierarchic structure of knowledge* is not only economical: we do not reproduce beside and for each "node" of dog in our memory (thousands of nodes) hundreds of data. We just write that around the class node and then - when needed - instantiate it on a specific dog.

The greatest advantage is not just in memory space and costs, but in the fact that we *know a lot of thing about something that we never met*; just by inference, prediction, inheritance. We have a lot of knowledge about a given entity without any direct experience of it. This crucial power of our cognitive organization is obviously exploited also in social life, in order to have information and expectations about people that we never met.

This fundamental device for "knowing without knowing" also about people is surely crucial also for *trust* evaluations and relations. Society works also on the basis of trust between strangers; this trust is based on several inferential and social tricks (like evoked feelings, analogy, recommendations, etc.) but is also strongly relying on *categories* of people and their "signaling" and recognition. If we (dis-)trust a given class of people and we understand that Y belongs to that class we can (dis-)trust Y.

The problem is:

- How do we build our trust in a category? From our direct experience or trust in its members? How many of them are necessary in order to generalize?
- How much risky is the instantiation from the class to that member Y? How much reliable are "signals" about Y membership? How much Y is representative, typical, of that class? And how much variance of trustworthiness there is in that class?
- When and how much it is advantageous to exploit trust on the categories and not just direct trust in the individual?

In this study we intend to explain and experimentally show the advantage of trust evaluation based on classes' reputation with respect to the reputation and opinion on single potential agents (partners). In an open world or in a broad population how can we have sufficient direct or reported experience on everybody? The quantity of potential agents in that population or net that might be excellent partners but that nobody knows enough can be high.

Our claim is that: the larger the population and the ignorance about the trustworthiness of each individual the more precious the role of trust in categories. If I know (through signals, marks, declaration, ...) the class of a given guy/agent I can have a reliable opinion of its trustworthiness derived from its class-membership.

It is clear that the advantages of such cognitive power provided by categories and prejudices do not only depend on recommendation and reputation about categories. We can personally build - by generalization - our evaluation of a given category from our direct experience with its members (this fact happens in our experiments for the agents that later have to propagate their recommendation about). However, in this simulation we have in the trustor (which has to decide whom rely on) only a prejudice based on recommendations about that category and not its personal experience.

Under a certain degree on direct experiences and circulation of recommendations, the performance of the evaluation based on classes will perform better; and in certain cases there will be no alternative at all: we do not have any evaluation on that individual, a part from its category; either we work on inferential instantiation of trustworthiness or we loose a lot of potential partners. This powerful inferential device has to be strongly present in WEB societies supported by MAS.

We simplify here the problem of the generalization process, of how to form judgment about groups, classes, etc. by putting aside for example inference from other classes (higher or sub); we build opinion (and then its transmission) about classes on the bases of experience with a number of subjects of a given class.

In this work we are also interested in showing the difference between localized and non-localized knowledge. A localized world is a world strongly influenced by the spatial distance between agents; a non-localized world is independent by the spatial distances, in which the communicative distance follows different routs with respect to the spatial distance.

The first approach reflects the traditional social way to exchange information, before the advent of virtual communities, where communication is constrained by spatial distance.

However, nowadays we also use another way to exchange information: the Web. Here we have access to a more complex net of users; our choice follows (and is influenced) by different communicative links to the information sources.

We are interested in analyzing the utility of categories in these two different contexts, trying to understand if and how they affect its performance.

Human society vs digital society

The defined **human society** is the represented by the classical social paradigm, in which agents do not use computers/internet for communications or any kind of relation.

With **digital society** we mean the digital world in which users communicate and relate by the means of computers/internet.

Given these two definitions, let's start analyzing the difference between the two worlds.

The first and immediate kind of difference regards physical distance: in the human society communication happens to be limited by distance; this is no more true in the digital society, in which one can communicate with the entire world, without moving from its desk.

The consequence of this two different kind of communication is that in the first one an agent/trustor will frequently communicate with the same agents (its neighbors), while in the second one the agent/trustor, as it is in the possibility to contact a lot of other agents, will communicate with more agents.

We will model these two kinds of behavior in the simulations, trying to understand how they affect categorial knowledge.

Related works

Differently from (Burnett et al, 2010) (Fang et al, 2012) (Sensoy et al, 2014), in this work we do not address the problem of learning categorical knowledge and we assume that the categorization process is objective. Similarly to (Burnett et al, 2013), we give agents the possibility to recommend categories and this is the key point of this work.

In (Messina et al, 2016) the authors propose the use of trust and reputation for producing trust evaluation. Focusing on reputation, they computed the reputation that a_i has of another node a_j, with respect to c – which is a category of service – as the average of all the recommendations received by the other nodes of the community, suitably weighted by their recommendation reliability in order to hinder malicious behaviors.

This is actually a good way to reproduce reliability, as the information reported by each node is properly weighted for the reliability that the node has on providing recommendation for that specific category of service. It is an important point as authors consider that the reliability of a node as a service provider and as a recommender are not the same.

I don't agree with the fact that this formula blocks malicious behavior. Suppose in fact that I have two malicious recommenders with a low reliability. If they state that a_j has a maximal reliability, I will completely believe them, regardless of the fact that I know they are malicious.

In the majority of the cases available in the literature, the concept of recommendation is used concerning recommender systems (Adomavicius and Tuzhilin, 2005). These ones can be realized using both past experience (content-based RS)(Lops et al, 2011) or collaborative filtering, in which the contribute of single agents/users is used to provide group recommendations to other agents/users.

Focusing on collaborative filtering, the concepts of similarity and trust are often exploited (together or separately) to determine which contributes are more important in the aggregation phase (Massa and Avesani, 2007) (Than and Han, 2014).

For instance, in (De Meo et al, 2015) authors provide a system able to recommend to users group that they could join in Online Social Network. Here it is introduced the concepts of compactness of a social group, defined as the weighted mean of the two dimensions of similarity and trust. Even in (Guo et al, 2015) authors present a clustering-based recommender system that exploits both similarity and trust, generating two different cluster views and combining them to obtain better results. Another example is (De Meo et al, 2011) where authors use information regarding social friendships in order to provide users with more accurate suggestions and rankings on items of their interest.

A classical decentralized approach is referral systems (Yolum, and Singh, 2003), where agents adaptively give referrals to one another.

Information sources come into play in FIRE (Huynh et al, 2006), a trust and reputation model that use them to produce a comprehensive assessment of an agent's likely performance. Here authors take into account open MAS, where agents continuously enter and leave the system. Specifically, FIRE exploits interaction trust, role-based trust, witness reputation, and certified reputation to provide trust metrics.

The described solutions are quite similar to our work, although we contextualized this problem to information sources. However we do not investigate recommendations with just the aim of suggesting a particular trustee, but also for inquiring categories' recommendations.

2. Recommendation and reputation: definitions

Let us consider a set of agents *Ag₁, ..., Agₙ* in a given world. We consider that each agent in this world could have trust relationships with anyone else. On the basis of these interactions the agents can evaluate the trust degree of their partners, so building their judgments about the trustworthiness of the agents with whom they interacted in the past.

The possibility to access to these judgments, through recommendations, is one of the main sources for trusting agents outside the circle of closer friends. Exactly for this reason recommendation and reputation are the more studied and diffused tools in the trust domain (Ramchurn et al, 2004).

We introduce

$$\mathrm{Re}\, c_{x,y,z}(\tau) \qquad (1)$$

where $x,y,z \in \{Ag_1, Ag_2, ..., Ag_n\}$, we call *D* the specific set of agents: $D \equiv \{Ag_1, Ag_2, ..., Ag_n\}$

and $0 \le \mathrm{Re}\, c_{x,y,z}(\tau) \le 1$

t, as established in the trust model of (Castelfranchi and Falcone, 2010), is the task on which the recommender *x* expresses the evaluation about *y*.

In words: $\mathrm{Re}\, c_{x,y,z}(\tau)$ is the value of *x*'s recommendation about *y* performing the task *t*, where *z* is the agent receiving this recommendation.

In this work, for sake of simplicity, we do not introduce any correlation/influence between the value of the recommendations and the kind of the agent receiving it: the value of the recommendation does not depend from the agent to whom it is communicated.

So (1) represents the basic expression for recommendation.

We can also define a more complex expression of recommendation, a sort of *average recommendation*:

$$\sum_{x=Ag_1}^{Ag_n} \mathrm{Re}c_{x,y,z}(\tau)/n \qquad (2)$$

in which all the agents in the defined set of agents express their individual recommendation on the agent *y* with respect the task *t* and the total value is divided by the number of agents.

We consider the expression (2) as the <u>reputation</u> of the agent y with respect to the task t in the set D.

Of course the reputation concept is more complex than the simplified version here introduced (Conte and Paolucci, 2002)(Sabater-Mir, 2003).

It is in fact the value that would emerge in the case in which we receive from each agent in the world its recommendation about y (considering each agent as equally reliable).

In the case in which an agent has to be recommended not only on one task but on a set of tasks (t_1 , ..., t_k), we could define instead of (1) and (2) the following expressions:

$$\sum_{i=1}^{k} \mathrm{Re}c_{x,y,z}(\tau_i)/k \qquad (3)$$

that represents the *x*'s recommendation about *y* performing the set of tasks (t_1,..., t_k), where *z* is the agent receiving this recommendation.

Imagine having to assign a meta-task (composed of a set of tasks) to just one of several agents. In this case the information given from the formula

(3) could be useful for selecting (given the *x*'s point of view) on average (with respect to the tasks) the more performative agent *y*.

$$\sum_{x=Ag_1}^{Agn} \sum_{i=1}^{k} \mathrm{Re}\, c_{x,y,z}(\tau_i)/nk \qquad (4)$$

that represents a sort *of average recommendation* from the set of agents in D, about *y* performing the set of tasks (*t*$_1$, ..., *t*$_k$). *We consider the expression (4) as the* <u>*reputation*</u> *of the agent y with respect the set of tasks (t*$_1$ *, ..., t*$_k$*), in the set D.*

Having to assign the meta-task proposed above, the information given from the formula (4) could be useful for selecting on average (with respect to both the tasks and the agents) the more performative agent *y*.

Using Categories

As described above, an interesting approach for evaluating agents is to classify them in specific categories already pre-judged/rated and as a consequence to do inherit to the agents the properties of their own categories.

So we can introduce also the *recommendations about categories*, not just about agents (we discuss elsewhere how these recommendations are formed). In this sense we define:

$$\mathrm{Re}\, c_{x,Cy,z}(\tau) \qquad (5)$$

where $x \in \{Ag_1, Ag_2, ..., Ag_n\}$ as usual, and we characterize the categories $\{C_1, ..., C_l\}$ through a set of features $\{f_{y1}, ..., f_{ym}\}$:

$$\forall y \in \{Ag_1, ..., Ag_n\} \exists c_y \in \{C_1, ..., C_l\} \mid (C_y \equiv \{f_{y1}, ..., f_{ym}\}) \wedge (\{f_{y1}, ..., f_{ym}\} \in y)$$

it is clear that there is a relationship between task *t,* and the features $\{f_{y1}, ..., f_{ym}\}$ of the C$_y$ category. In words we can say that each agent in *D* is classified in one of the categories $\{C_1, ..., C_l\}$ that are characterized from a set of features $\{f_1,...,f_m\}$; as a consequence each agent belonging to a category owns the features of that category. $0 \le \mathrm{Re}c_{x,Cy,z}(\tau) \le 1$

In words: $\mathrm{Re}c_{x,Cy,z}(\tau)$ *is the value of x's recommendation about the agents included in category* C_y *when they perform the task t,* (as usual *z* is the agent receiving this recommendation).

We again define a more complex expression of recommendation, a sort of *average recommendation*:

$$\sum_{x=Ag_1}^{Ag_n} \mathrm{Re}c_{x,Cy,z}(\tau)/n \qquad (6)$$

in which all the agents in the domain express their individual recommendation on the category C_y with respect the task *t* and the total value is divided by the number of the recommenders.

We consider the expression (6) as the <u>reputation</u> of the category C_y *with respect the task t in the set D.*

Now we extend to the categories, in particular to C_y, the recommendations on a set of tasks (t_1, ...,t_k):

$$\sum_{i=1}^{k} \mathrm{Re}c_{x,Cy,z}(\tau_i)/k \qquad (7)$$

that represents *the recommendation value of the x's agent about the agents belonging to the category* C_y *when they perform the set of tasks* (t_1,...,t_k).

Finally, we define:

$$\sum_{x=Ag_1}^{Agn} \sum_{i=1}^{k} \mathrm{Re}c_{x,Cy,z}(\tau_i)/nk \qquad (8)$$

that represents *the value of the reputation of the category* C_y *(of all the agents y included in* C_y*) with respect the set of tasks* (t_1,...,t_k)*, in the set D.*

3. Definition of Interest for this Work

In this work we are in particular interested in the case in which *z* (a new agent introduced in the world) asks for recommendation to *x* ($x \in D$) about an agent belonging to its domain *D*$_x$ for performing the task *t* (*D*$_x$ is a subset of *D*, it is composed by the agents that *x* knows). *x* will select the best evaluated *y*, with $y \in D_x$ on the basis of formula:

$$\max_{y \in D_x}(\mathrm{Re}\,c_{x,y,z}(\tau)) \qquad (9)$$

where $D_x \equiv \{Ag_1, Ag_2, ..., Ag_m\}$, D_x includes all the agents evaluated by *x*. They are a subset of *D*: $D_x \subseteq D$.

In general *D* and *D*$_x$ are different because *x* does not necessarily know (has interacted with) all the agents in *D*.

z asks for recommendations not only to one agent, but to a set of different agents: $x \in D_z$ (*D*$_Z$ is a subset of *D*, to which z asks for reputation), and selects the best one on the basis of the value given from the formula:

$$\max_{x \in D_z}(\max_{y \in D_x}(\mathrm{Re}\,c_{x,y,z}(\tau))) \qquad (10)$$

$D_z \subseteq D$, z could ask to all the agents in the world or to a defined subset of it (see later).

We are also interested to the case in which *z* ask for recommendations to *x* about a specific *agents' category* for performing the task *t*. *x* has to select the best evaluated C$_y$ among the different $C_y \in \{C_1, ..., C_l\}$ x has interacted with (we are supposing that each agent in the world *D*, belongs to a category in the set $\{C_1, ..., C_l\}$).

In this case we have the following formulas:

$$\max_{Cy \in D_x}(\mathrm{Re}\,c_{x,Cy,z}(\tau)) \qquad (11)$$

that returns the category best evaluated from the point of view of an agent (x). And

$$\max_{x \in D_z}(\max_{Cy \in D_x}(\mathrm{Re}\, c_{x,Cy,z}(\tau))) \qquad (12)$$

that returns the category best evaluated from the point of view of all the agents included in D_z.

4. Computational Model

In order to realize our simulations, we exploited the software NetLogo (Wilensky, 1999).

In every scenario there are four general categories, called Cat1, Cat2, Cat3 and Cat4, composed by 100 agents per category. Each category is characterized by:

1. an **average value of trustworthiness**, in range [0,100];
2. an **uncertainty value**, in range [0,100]; this value represents the interval of trustworthiness in which the agents can be considered as belonging to that category.

These two values are exploited to generate the **objective trustworthiness** of each agent, defined as *the probability that, concerning a specific kind of required information, the agent will communicate the right information*.

Of course the trustworthiness of categories and agents is strongly related to the kind of requested information/task. Nevertheless, for the purpose of our it is enough to use just one kind of information (defined by *t*) in the simulations. The categories' trustworthiness of Cat1, Cat2, Cat3 and Cat4 are fixed respectively to 80, 60, 40 and 20% for *t*. What changes through scenarios is the uncertainty value of the categories: 1, 20, 50, and 80%.

We want to present a series of scenarios with different settings and referred to **human and digital worlds**, to show when it is more convenient to exploit recommendations about categories rather than recommendations about individuals, and vice versa.

Both the simulations are composed by two main steps that are repeated continuously. In the first step, called **exploration phase**, agents without any knowledge about the world start experiencing other agents, asking to a subset of the population for the information P. Then they memorize the performance of each queried agent both as individual element and as a member of its own category.

The performance of a agent can assume just the two values 1 or 0, with 1 meaning that the agent is supporting the information P and 0 meaning that it is opposing to P. For sake of simplicity, we assume that P is always true.

The exploration phase has a variable duration, going from 100 ticks to 1 tick. Depending on this value, agents will have a better or worse knowledge of the other agents.

Then, in a second step (**querying phase**) we introduce in the world a trustor (a new agent with no knowlegde about the trustworthiness of other agents and categories, and that has the necessity to trust someone reliable for a given informative task: in our case t). It will select a given subset of the population and it will query them. In particular, the trustor will ask them for the best category and the best trustee they have experienced.

In this way, the trustor is able to collect information about both the best recommended category and agent.

It is important to underline that the trustor is collecting information from the agents considering them as equally trustworthy with respect to the task of "providing recommendations". Otherwise it should weigh differently these recommendations. In practice our agents are sincere.

Then it will select an agent belonging to the best recommended category and it will compare it, in terms of objective trustworthiness, with the best recommended individual agent (trustee).

The possible **outcomes** are:

- **trustee wins (t_win)**: the trustee selected with individual recommendation is better than the one selected by the means of category; then this method gets one point;
- **category wins (c_win)**: the trustee selected by the means of category is better than the one selected with individual recommendation; then this method gets one point;
- **equivalent result**: if the difference between the two trustworthiness values is not enough (it is under a threshold), we consider it as indistinguishable result. In particular, we considered the threshold of 3% as, on the basis of previous test simulations, it has resulted a reasonable value.

These two phases are repeated 500 times for each setting.
In particular, we will represent this value:

$$\frac{c\_win}{c\_win + t\_win} \qquad (13)$$

This ratio shows how much categories' recommendation is useful if compared to individual recommendation.
Simulations' results are presented in a graphical way, exploiting 3D shapes to represent all the outcomes. These shapes are divided into two area and represented with two different colors:

- the part over 0.5, represented in light gray, in which prevails the category recommendation;
- the one below 0.5, represented in dark gray, in which prevails the individual recommendation.

These graphs represent a useful view about the utility of the categorial role in the different interactional and social contexts.

For each value of uncertainty, we explored 40 different settings, considering all the possible couple of **exploration phase** and **queried trustee percentage**, where:

- exploration phase ∈ {all-in,1,3,5,10,25,50,100};
- queried trustees' percentage ∈ {5,10,25,50,100}.

When the exploration phase assume the value "all-in" the exploration lasts just 1 tick and in that tick every agent experiences all the others. Although this is a limit case, very unlikely in the real world, it is really interesting as each agent has not a good knowledge of the other agent as individual elements (it has experienced them just one time), but it is able to get a really good knowledge of their categories, as it has experienced them as many times as the number of agents for each category. So this is an explicit case in which the recommendations of the agents about categories are surely more informative than the ones about individuals.

First simulation: human society

As previously said, in this first simulation everything is ruled by physical distance:

1. in the exploration phase, on each tick agents move into the world with a probability of 10%; this has the purpose of creating a localization phenomena; then agents will ask for information P to the other trustees which distance in less than 3 NetLogo patches; empirically, we saw that on average they select the 3% of the population (in order to be similar to the second simulation);

2. in the querying phase, given a percentage of population going from 100% to 5%, the trustor will select the first neighbors until it reaches the requested percentage;
3. in the end, the trustor will select the nearest member of the most recommended category, to compare it with the most recommended agent.

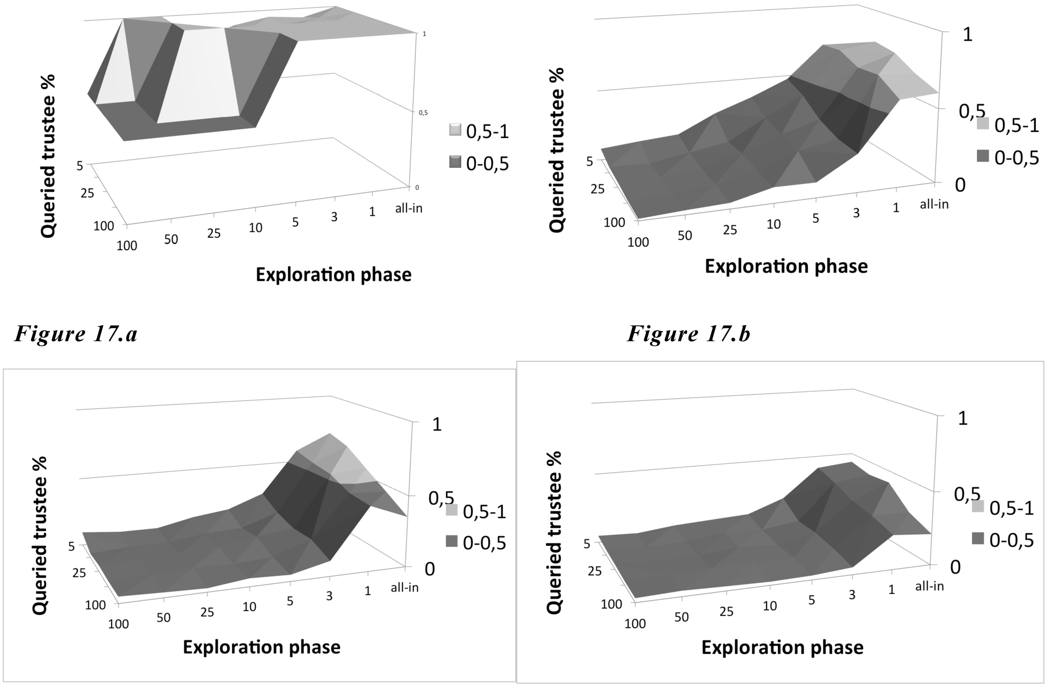

Figure 17.a

Figure 17.b

Figure 17.c

Figure 17.d

Figure 17.a, 17.b, 17.c and 17.d stand respectively for 1%, 20%, 50% and 80% of categories' uncertainty

Second simulation: digital society

Conversely from the previous one, in this simulation we explore the case in which the communication in the world is not limited by the phisical distance, like in the web context.

Here we will have that:

1. concerning the exploration phase, agents will ask for information P to a random 3% of the poputalion;
2. concerning the querying phase, the trustor will select (again in a random way) a given subset of the population, going from 100% to 5%;
3. in the end, the trustor will select a random member of the most recommended category, to compare it with the most recommended agent.

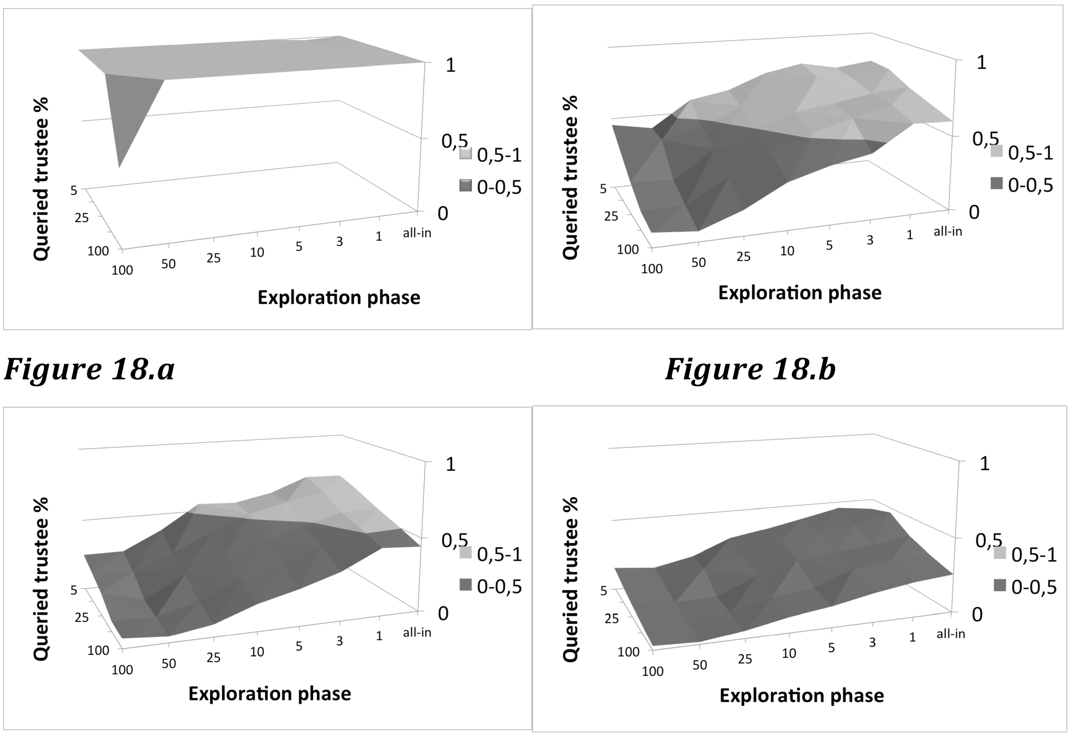

Figure 18.a

Figure 18.b

Figure 18.c

Figure 18.d

Figure 18.a, 2.b, 2.c and **2.d** stand respectively for 1%, 20%, 50% and 80% of categories' uncertainty

Results' discussion

Effects in each simulation

Let's start analyzing the parameters that influences both the two simulations. In particular, we identify three effects that influence the outcome. The **first effect** is due to categories' uncertainty: the less it is, the more is the utility of using categories; the more it is, the less categories will be useful. It is not possible to notice this effect just looking at one picture. On the contrary, looking at the overall picture one can notice that the curves of the graphs lower, going from a maximal value in **Figure 1.a** and **2.a** to a minimal value in **Figure 1.d** and **2.d**.

The **second effect** is due to exploration phase. The longer this phase is the more individual recommendations are useful; the less it lasts the more category recommendations are useful.

The **third effect** is introduced by the queried trustee percentage, that acts exactly as the exploration phase: the higher the percentage of queried agents, the more individual's recommendations are useful; the less it is, the more categories' recommendations are useful.

The length of the exploration phase and the percentage of the queried agents occur in all of the four graphs and cooperate in determining respectively the degree of knowledge (or ignorance) in the world and the level of inquire about this knowledge. In particular, with "the knowledge in the world" we intend how the agents can witness the trustworthiness of the other agents or their aggregate, given the constraints defined from the external circumstances (number and kind of interactions, kind of categories, and so on).

In practice, both these elements seem to suggest how the role of categories becomes relevant when either decreases or degrades the knowledge within the analyzed system (before the interaction with the trustor) or is reduced the transferred knowledge (to the trustor).

Let us explain better. The *first effect* shows how the reliability of category's trustworthiness (that will be inherited by its members) depends, of course, from the variability of the behavior among the class members. There may be classes where all the members are very correct and competent, other classes where there is a very high variance: in this last case our betting on a member of that class is quite risky.

The *second effect* can be described with the fact that each agent, reducing the number of interactions with the other agents in the explorative phase, will have relevantly less information with respect to the individual agents. At the same time its knowledge with respect to categories does not undergo a significant decline given that categories' performances derive from several different agents.

The *third effect* can be explained with the fact that reducing the number of queried trustees, the trustor will receive with decreasing probability information about the more trustworthy individual agents in the domain, while information on categories, maintains a good level of stability also reducing the number of queried agents, thanks to greater robustness of these structures.

Resuming, the above pictures clearly show how, when the quantity of information (about the agents' trustworthiness exchanged in the system) decreases, it is better to rely on the categorial recommendations rather than individual recommendations.

This result reaches the point of highest criticality in the “all-in” case in which, as expected, the relevance of categories reaches its maximal value.

Human Society and Digital Society

Let’s then discuss the difference between these two main setting: the human society (HS) and the digital society (DS).

The *first difference* between them is in the behavior. In fact, while the DS tends to have a convex behavior, the HS one tends to be concave: the descent of the categories’ utility in the first case is less steep than in the second.

The *second effect* is easier to notice: the curves of DS case are quite always higher than the HS case.

Both these effects are symptoms of the fact that the utility of categories is higher in the DS case.

In fact, in the DS the agents can have access to more other agents, as they are not constrained by physical distance. In this way, they know more agents, but their knowledge about each single agent is limited.

Conversely in the HS, each agent can ask just to its neighbors. Although they move into the world, their knowledge is strictly related to their physical position. As a consequence, they will know better their neighbors and their knowledge of categories strongly depends on the individuals they have met.

5. Conclusion

Other works (Burnett et al, 2010) (Castelfranchi and Falcone, 2010) show the advantages of using reasoning about categorization for selecting trustworthy agents. In particular, how it was possible to attribute to a certain unknown agent, a value of trustworthiness with respect to a specific

task, on the basis of its classification in, and membership to, one (/or more) category/ies. In practice, the role of generalized knowledge has proven to determine the possibility to anticipate the value of unknown agents.
In this work we investigated the different roles that recommendations about individual agents and about categories of agents can play, both in human society and digital society.

We also showed cases in which information about categories is more useful than information towards individual agents, inquiring and matching different dimensions and parameters. This kind of analysis can be particularly relevant to decide how to built the cognitive approach of agents searching information among multiple sources. Before choosing between direct or generalized information, we need to evaluate how information is distributed among the agents in the specific domain. Our results show that in certain cases becomes essential the use of categorial knowledge for selecting qualified partners.

In particular we showed how the categorial knowledge becomes critical in the new digital society, in which one interacts with an higher number of agents, so that the dimension past experience loses importance, letting the categorial one take over.

CHAPTER 5:

EXPLOITING INFORMATION SOURCE IN CASE OF CRITICAL HYDROGEOLOGICAL RISK

In this work we present a study about cognitive agents that have to learn how their different information sources can be more or less trustworthy in different situations and with respect to different hydrogeological phenomena. We introduced the realized platform that can be manipulated in order to shape many possible scenarios.

The simulations are populated by a number of agents that have *three information sources* about forecasts of different hydrogeological phenomena. These sources are: a) their own evaluation/forecast about the hydrogeological event; b) the information about the event communicated by an authority; the behavior of other agents as evidence for evaluating the dangerous level of the coming hydrogeological event.

These weather forecasts are essential for the agents in order to deal with different and more or less dangerous meteorological events requiring adequate behaviors. We consider in particular in this work some specific situations in which the authority can be more or less trustworthy and more or less able to deliver its own forecasts to the agents. The simulations will show how, on the basis of a training phase in these different situations, the agents will be able to make a rational use of their different information sources.

This work has resulted in the following publications:

1. Falcone, R., Sapienza, A., Castelfanchi, C., Information Sources about Hydrological Disasters: The Role of Trust, in proceedings of EUMAS 2015, LNAI 9571, pp. 350-362, 2015, Springer.
2. Sapienza, A., Falcone, R., "How manage the information sources' trustworthiness in a scenario of hydrogeological risks", in proceedings of TRUST 2016 in conjunction with AAMAS 2016, Singapore, ceur-ws, vol 1578, paper 11.
3. Falcone R., Sapienza A. and Castelfranchi C. (2016). Trusting Different Information Sources in a Weather Scenario: A Platform for Computational Simulation.In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART,* ISBN 978-989-758-172-4, pages 165-172. DOI: 10.5220/0005695501650172
4. Falcone, R., Sapienza, A., & Castelfranchi, C. (2016). Which information sources are more trustworthy in a scenario of hydrogeological risks: a computational platform. In *Advances in Practical Applications of Scalable Multi-agent Systems. The PAAMS Collection* (pp. 84-96). Springer International Publishing.
5. Falcone, R., & Sapienza, A. (2016). An Evolutionary Platform for Social Simulation in Case of Critical Hydrogeological Phenomena: The Authority's Role. In *Advances in Practical Applications of Scalable Multi-agent Systems. The PAAMS Collection* (pp. 251-255). Springer International Publishing.

1. Introduction

One of the main problems we have for understanding and foreseeing any domain and world is not just to access to the different information sources about that domain, but also to be able to evaluate the trustworthiness of those sources.

In particular, the same source not necessarily has the same degree of trustworthiness in any situation and context, but could change its own reliability on the basis of different external or internal factors. For example, in the domain of the weather forecasts, we know that the mathematical models used for this task are quite reliable when referring to temporally close events (next 6-24 hours), while they are quite approximate if referring long-term events (1-4 weeks). And we also know that these forecasts can change their reliability on the basis of the kind of phenomenon they are evaluating.

So it can be very relevant to have different information sources and also to know how trustworthy they are in different contexts and situations.
On the other hand, trying to put together information coming from different information sources can be an uneasy task. It is necessary to have strategies to do it, especially in presence of critical situation, when there are temporal limits to get decision and a wrong choice can lead to an economical loss or even to risk life.

As said, the necessity of integrating sources on different scopes can be very useful in order to make a well-informed decision. In case of the weather forecast we can consider different sources: official bulletin of authorities, the observation of other agents' behavior and of their decisions during the meteorological event, the direct evaluation and competence of the same agents as the basis for their own decisions.

Some of these sources are not correlated among them: a forecast is referred to mathematical model of the weather linked to its previous data, while a direct evaluation can be based on a current human perception of the phenomenon (with its potential psychological and perceptive bias). Then, integrating these sources becomes essential and at the same time it is necessary to identify and take into account their trustworthiness.

Again, according to our view trusting an information source (S) means to use a cognitive model based on the dimensions of competence and reliability/motivation of the source. These competence and reliability evaluations can derive from different reasons, basically:

- Our previous *direct experience* with S on that specific kind of information content.
- *Recommendations* (other individuals Z reporting their direct experience and evaluation about S) or *Reputation* (the shared general opinion of others about S) on that specific information content;
- *Categorization* of S (it is assumed that a source can be categorized and that it is known this category), exploiting inference and reasoning (analogy, inheritance, etc.).

However in this work, for sake of simplicity, we use just the first kind of the three reasons above described: the direct experience with each source.

Our agents manipulate their values of trust (we consider the feedback effects and the trust dynamics). In practice each agent evaluates if an information source was corrected with its own prediction and, on the basis of this evaluation, decides to increase its trustworthiness or decrease it.

When we start the simulations we have a number of agents that equally trust their three different sources (neutral agents). Then we let them make experience with their sources by the means of a training period, using different weather scenarios. We do this with the presence of four kinds of authorities: reliable and strongly communicative, reliable and weakly communicative, not reliable and strongly communicative, and not reliable and weakly communicative.

Our investigation is about: are the agents able to learn the more trustworthy sources? Are they able to intelligently integrate these sources? Are the agents' performances coherent with the trustworthiness of the sources they are following? Are we able to extract useful information from these simulations for situations of real cases? In the work we show how, with a certain limit of approximation, we can give some useful and interesting indications.

2. The Trust Model

In this work we use the cognitive model for trust on information sources proposed in the section "**THE MODEL WITH MULTIPLE SOURCE ASSERTING (OR DENYING) A SINGLE BELIEF**" of chapter 2.

However in these work trust is not just a static value but it is used in a dynamic way. In order to do that, we added to the computational model the concept of feedback on trust.

Feedback on trust

Trust is a dynamic concept. It changes with time because the situation can change. Someone that once was competent in a given subject or topic can improve its ability (or on the contrary it could loose it). Maybe it is no more motivated to help us or it is interested in deceiving us.

Given that, it emerges the necessity for the agents to adapt to the context in which they move.

In our simulations we want that, starting from a neutral trust level (that does not imply trust or distrust), agents will try to understand how much to rely on each single information source.

To do that, they need a way to perform feedback on trust. They need to evaluate the current performance of their trustees/information sources so that they can adjust they trust they have in them.

The simple but effective way to do it is to exploit weighted mean. This method just need to establish how much weight the new performance should have for the new evaluation.

Given the two parameters α and β, the new trust value is computed as:

$$newTrustDegree = \alpha * oldTrustDegree + \beta * performanceEvaluation$$

$$\alpha + \beta = 1$$

Of course changing the values of α and β will have an impact on the trust evaluations. With high values of α/β, agents will need more time to get a precise evaluation, but a low value (below 1) will lead to an unstable evaluation, as it would depend too much on the last performance. We do not investigate these two parameters in this work, using respectively the values 0.9 and 0.1. In order to have good evaluations, we let agents make a lot of experience with their information sources.

In the formula, $oldTrustDegree$ is the previous trust degree and the $performanceEvaluation$ is the objective evaluation of the source performance. This last value is obtained comparing what the source said with what actually happened. Considering the PDF reported by the source, and remembering that it is split into three parts, we will have that:

1. The estimated probability of the event that actually occurred is completely taken into account;
2. The estimated probability of the event immediately near to the actual one is taken into account for just 1/3.
3. The rest of the PDF is not considered.

It is worth underlining that the process of feedback on trust takes the PDF as input, not the SPDF. We need to consider what actually the source reported as it reported, while in SPDF information has been filtered due to the personal consideration of the trustor.

Let's show an example to explain better how this process works. Suppose that there has been a critical event. A first source reported in its PDF a 100% probability of critical event; a second one a 50% probability of critical event and a 50% of medium event; finally a third one asserts 100% of light event

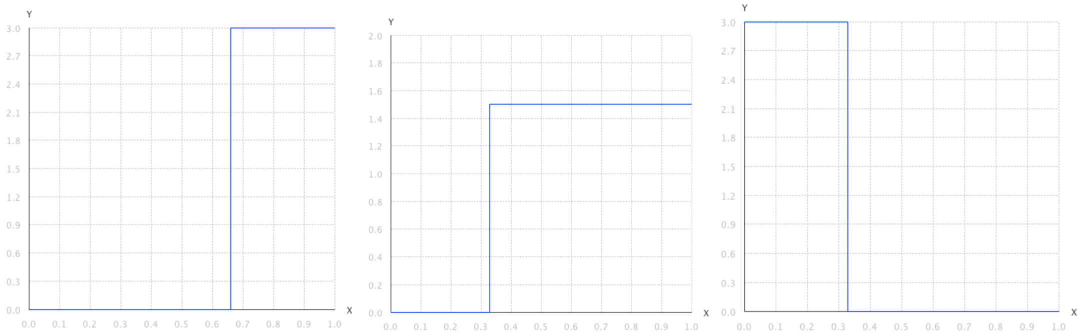

Figure 19. (a) A source reporting a 100% probability of critical event. (b) A source reporting a 50% probability of critical event and 50% probability of medium event. (c) A source reporting a 100% probability of light event

As the first source provided a completely correct PDF, its evaluation will be 100%.

The third source reported a 100% probability of light event, therefore the corresponding PDF is completely wrong. Its evaluation will be 0%.

The second case is the most interesting one. Here we have that the source is unsure about the medium event and the critical event. As there has been a critical event, this part is considered as correct (50%). Concerning the 50% probability of medium event, we state that even if it is not correct, it is not even completely wrong. Then we will consider 1/3 of it. The final evaluation for this performance will be 50% + (50/3)% = 66.67%

This is an important point. We in fact suppose that there is a difference inside wrong information. As the world is not Boolean (true/false, correct/wrong and so on) the correctness of information can be considered with different level. We cannot just say that if information is not correct it is wrong: there are different degrees of correctness! In this case saying that there will be a light event is more wrong that saying that there will be a medium event. It is necessary to understand how much information given by the source is distant by the correct information. This is the reason why we decide to model this way the evaluation of a source’s performance.

3. THE PLATFORM

Exploiting NetLogo (Wilensky, 1999), we created a very flexible platform, where a lot of parameters are taken into account to model a variety of situations.

The Context

The basic idea is that, given a population distributed over a wide area, some weather phenomena happen in the world with a variable level of criticality. These weather phenomena happen in the world in a temporal window of 16 ticks.
The world is populated by a number of cognitive agents (citizens) that react to these situations, deciding how to behave, on the basis of the information sources they have and of the trustworthiness they attribute to these different sources: they can escape, take measures or evaluate absence of danger.

In addition to citizens, there is another agent called authority. Its aim is to inform promptly citizens about the weather phenomena. The authority will be characterized by an uncertainty, expressed in terms of standard deviation, and by a communicativeness value, that represents the probability that it will be able to inform each single citizen.

Information Sources

To make a decision, each agent can consult a set of information sources, reporting to it some evidence about the incoming meteorological phenomena.

We considered the presence of three kinds of information sources (whether active or passive) for agents:

1. Their *personal judgment*, based on the direct observation of the phenomena. Although this is a direct and always true (at least in that moment) source, it has the drawback that waiting to see what happens could lead into a situation in which it is no more possible to react in the best way (for example there is no more time to escape if one realizes too late the worsening weather).
2. *Notification from authority*: the authority distributes into the world weather forecast with associated different alarm signals, trying to prepare citizens to what is going to happen. It is not sure that the authority will be able to inform everyone
3. *Others' behavior*: agents are in some way influenced by community logics, tending to partially or totally emulate their neighbors behavior.

The notification from the authority is provided as a clear signal: all the probability is focused on a single event.

Conversely, the personal judgment can be distributed on two or three events with different probabilities. This can also be true for others' behavior estimation as the probability of each event is directly proportional to the number of neighbors making each kind of decision. If no decision is available, the PDF is a uniform distribution with value 1.

Agents' Description

At the beginning of the simulation, the world is populated by a number of agents. These agents have the same neutral trust value 0.5 for all their information sources. This value represents a situation in which agents are not sure if to trust or not a given source, as a value of 1 represents complete trust and 0 stands for complete distrust.

Agents are also characterized by a decision deadline, expressed in ticks, that determines the moment in which agents will make a decision.

The unique difference between them relies on how much they are able to see and to read the phenomena. In fact, in the real world not all the agents have the same abilities. In order to shape this, we divided equally agents into three set:

1. Good evaluators: agents that are able to see 15 ticks of event. They will be quit always able to detect correctly the event, then we expect them to rely mainly on their own opinion.
2. Medium evaluators: agents that are able to see 14 ticks of event. They can detect the event, but not good as the previous category.
3. Bad evaluators: agents that are able to see 13 ticks of event. Quite often, they will detect two possible events, but they will need another source to decide between them.

World Description

The world is made by 32x32 patches, which wraps both horizontally and vertically. It is geographically divided in 4 quadrants of equal dimension, where agents are distributed in a random way.

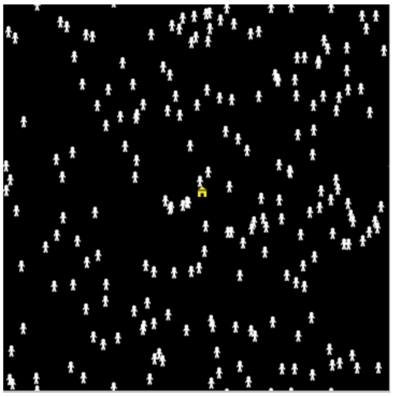

Figure 20. A world example. There are 200 agents plus the authority, represented by the yellow house in the center.

The quadrants differ in the possible weather phenomena that happen, modeled through the presence of clouds. The events are modeled so that agents can't be completely sure of what is going to happens:

1. *Critical event*: a tremendous event due to a very high level of rain, with possible risks for the agents sake; it is presented through a 16 ticks sequence of 3 clouds;
2. *Medium event*: it can cause possible damage to house or streets, but there is not health hazard; it is composed by a 16 ticks sequence of 2 or 3 clouds (or by a sequence of 13 ticks at the beginning followed by at least a couple of (2,3) in any sequence). To let this event be similar to the critical one, the 50% of the times we force the firsts 13 ticks to be equal to 3.
3. *Light event*: there is not enough rain to make any damage. It is composed by a 13 ticks sequence of 2/3 ticks followed by 2 ticks that can assume one of the value {0,1,2,3} and then a 0. Then, this event can be confused with a medium one (if not seen in its completeness).

As seen, these phenomena are not instantaneous, but they happen progressively in time, adding a given number of clouds on each tick until the phenomenon is completed.

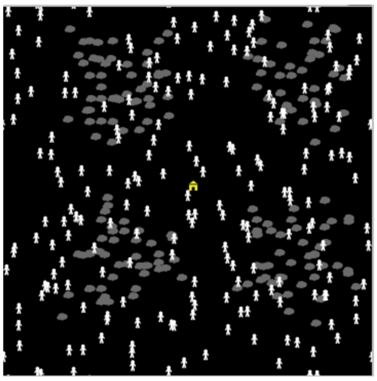

Figure 21. An example of a world after an event. Starting from the quadrants in the upper left and proceeding clockwise, we can see events 3, 1, 3 and 2.

The four quadrants are independent from each other but there can be an indirect influence as agents can have neighbors in other quadrants.
In each quadrant, each event has a fixed probability to happen:

1. 10% for critical event;
2. 20% for medium event;
3. 70% for light event.

These events are also correlated to the alarms that the authority raises. In fact, as previously said, the authority is characterized by a standard deviation. We use it to produce the alarm generated by the authority and from it depends the correctness of the prediction. In particular, we considered four kinds of authorities: **reliable and strongly communicative**, **reliable and weakly communicative**, **non reliable and strongly communicative**, **non reliable and weakly communicative**.

Own Evaluation

How should agents evaluate the phenomena they see? We propose an empirical way to evaluate them, taking into account how phenomena are generated and how they can evolve.

Considering what we just said, agents can see a sequence of 3 clouds or a sequence of 2 and 3 clouds. The first one can lead to a critical or a medium event, the second one to a medium or light event.

Table 20 shows a complete decryption of how agents evaluate what they see:

Table 20. This table provides a comprehensive description of agents' evaluation in each possible situation

Own Evaluation	Own Evaluation	Own Evaluation	Own Evaluation
15 ticks of 3	90%	10%	0%
14 ticks of 3	80%	20%	0%
13 ticks of 3	50%	50%	0%
12 ticks of 3	10%	80%	10%
11 ticks of 3	5%	70%	25%
13 ticks of 3 and 1 tick of 2	0%	80%	20%
13 ticks of 3 followed by (2,2), (2,3) or (3,2)	0%	100% (full-blowen) medium event	0%
15 ticks of 2 or 3	0%	90%	10%
14 ticks of 2 or 3	0%	80%	20%
13 ticks of 2 or 3	0%	50%	50%
12 ticks of 2 or 3	0%	20%	80%

11 ticks of 2 or 3	0%	10%	90%
Any other case	0%	0%	100%

Programmatically, agents make a pattern matching of what they see, respecting the order of the table. As readers can notice, there are just a few cases in which agents are completely sure of what is going to happen.

Workflow

Each simulation is divided into two steps. The first one is called "**training phase**" and has the aim of letting agents make experience with their information sources, so that they can determine how much each source is reliable.

At the beginning of this phase, we generate a world containing an authority and a given number of agents, with different abilities in understanding weather phenomena.

At the time t_0 the authority gives forecast for a future temporal window (composed by 16 ticks) including an alarm signal, reporting the level of criticality of the event that is going to happen in each quadrant (critic = 3, medium = 2, light =1). This information will reach each single agent with a probability given by the **authority communicativeness**.

Being just a forecast, it is not sure that it is really going to happen. It will have a probability linked to the precision of the authority (depending from standard deviation). However, as a forecast, it allows agents to evaluate the situation in advance, before the possible event. Event in fact starts at t_1 and, as previously said, lasts for 16 ticks.

During the decision making phase, agents check their own information sources, aggregating the single contributes according to the corresponding trust values. They estimate the possibility that each event happens and take the choice that minimizes the risk. Then, accordingly to their own decision-making deadlines, agents will choose how to behave.

While agents collect information they are considered as “thinking”, meaning that they have not decided yet. When this phase reaches the deadline, agents have to make a decision, that cannot be changed anymore. This information is then available for the other agents (neighborhood), which can in turn exploit it for their decisions.

At the end of the event, agents evaluate the performance of the source they used and adjust the corresponding trust values. If they haven’t been reached by the authority, there will not be a feedback on trust but, as this source wasn’t available when necessary, there will be a reduction of trust linked to the kind of event that happened: -0.15 for a critical event, -0.1 for a medium event, -0.05 for a light event.

This phase is repeated for 100 times (then there will be 100 events) so that agents can make enough experience to judge their sources.

After that, there is the “**testing phase**”. Here we want to understand how agents perform, once they know how much reliable their source are. In order to do that, we investigate how they perform in presence of a fixed map [3 1 3 2]. In this phase, we will compute the accuracy of their decision (1 if correct, 0 if wrong).

The decision-making phase

Once consulted all the three sources of information, agents subjectively estimate the probability that each single event happens:

1. $P_{critical\_event}$ = probability that there will be a critical event;
2. P_{medium_event} = probability that there will be a medium event;

3. P_{light_event} = probability that there is a light event;

They will react according to the event that is considered more likely to happen.

There are three possible choices:

1. Escape: agents abandon their homes.
2. Take measures: agents take some measure (quick repairs) to avoid possible damages due to weather event;
3. Ignore the problem: agents continue doing their activities, regardless of possible risks.

We assume that there is a time limit for taking a decision. This deadline is fixed to 15 ticks. Agents have to decide within this moment.

Platform Input

The first thing that can be customized is the **number of agents** in the world. Then, one can set the value of the two parameters **α and β**, used for the sources' trust evaluation.

It is possible to change the **authority reliability**, modifying its standard deviation, and the **authority communicativeness**, that represents the probability that each single citizen will receive the message of the authority.

Concerning the training phase, it is possible to change its **duration** and determine the **probability of the events** that are going to happen on each quadrant, while in the testing phase, that lasts just for 1 event, one can configure what we call the **event map**: it is the set of the four events relative to the four quadrants, starting from the one top left and proceeding clockwise.

4. SIMULATIONS

Once realized the platform, we decided to use it to investigate how different authority's behavior affects citizens choice and the trust they have in their information sources. We believe in fact that authority's choices affect not only citizens individually and directly, but also, by a social effect, citizens not directly reached by the authority. To do that, we investigated a series of scenarios, populated by equal populations, but in presence of different authorities. Then we analyze how citizens respond to these changes, measuring their trust values and the choice they make in presence of possible risks.

Simulations results are mediated through 500 cases, in order to delete the variability that a single case can have.

Simulation's results

Simulation settings:

1. **number of agents:** 200;
2. **α and β**: respectively 0.9 and 0.1;
3. **authority reliability**: we used the value 0.3 to shape a very reliable authority (its forecast are correct about the 90% of time) and 0.9 to shape a non reliable authority (its forecast are correct about the 50% of time);
4. **authority communicativeness**: 100% for all the three events (strongly communicative), meaning that each agent will receive authority's forecast or 30% for all the three events (weakly communicative), meaning that each agent will receive authority's forecast only the 30% of the time;
5. **training phase duration**: 100 events;

6. **probability of the events**: 10% critical event, 20% medium event, 70% light event;
7. **event map:** [3 1 3 2].

To help understanding our results, we are going to show them together. Let's start from the trust analysis

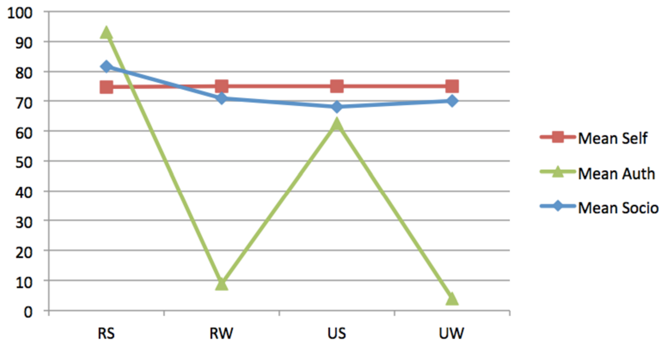

Figure 22: Here we represent the average trust value of all the agents on the three information sources in the four scenarios presented before.

Where RS = reliable strongly communicative, RW = reliable weakly communicative, US = unreliable strongly communicative, UW = unreliable weakly communicative.

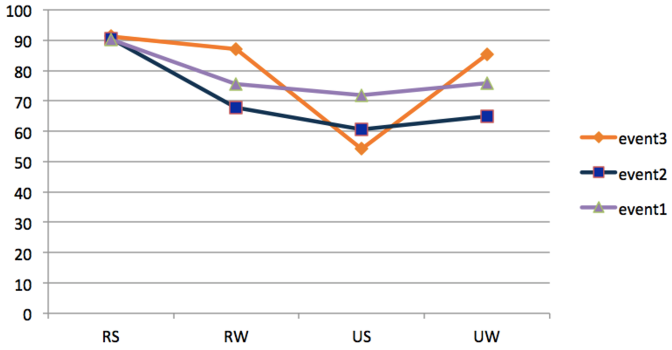

Figure 23: Here we represent the average performances of all the agents for the three events in presence of four different authorities.

Notice that, as implemented, 1/3 of citizens will quite always be able to quite completely understand the event, another 1/3 of them can understand quite well what is going to happen but it has less confidence on its evaluations, the remaining 1/3 of citizens is not a good evaluator. This fact stands for all the scenarios and that's why we have a standard value of average self trust.

Let's start analyzing results case by case.

In this first **case RS** (reliable strongly communicative), the authority is reliable and all the agents have access to its information. This leads to a high level of authority trust, but also to a high level of social trust, as the information communicated at the social level is due to the one coming from the authority and the one directly seen by citizens.

In the **RW case** (reliable weakly communicative) we have a reliable authority, but it rarely communicates its information. This involves a lack of trust in that source as, even if it is reliable, it is quite always not available. Socially, this results in a lower lever of social trust, as agents' decisions are just based on what they see (and just a part of them is able to directly understand the weather phenomenon). Even the agents' accuracy lowers down. Being the authority quite always unavailable, agents have to rely on their own abilities. In fact this effect is particularly strong for light events, as agents have less probability to read it correctly. Conversely, it is less strong on critical events, as is it easier to predict.

Let's see what happens when the authority is no more reliable, but it starts communicating wrong information. In the **US case** (unreliable strongly communicative), the average authority trust is 62.75%; this value comes from the fact that the authority reports correct information about the 50% of time. Remember that it is not true that wrong information is evaluated as

0. If it is not completely wrong (the authority predicted an event immediately near to the actual one) the evaluation will be 0.33.

Most of the agents identify their own observation as the better information source (as the authority in not so much trustworthy). Then social decision are mainly influenced by this component, but there is also a minimal influenced of the authority performance. That is why we have a decrement on social trust.

The last case, the **UW** (unreliable weakly communicative), is supposed to be the worst one; the authority is not reliable and is weakly communicative. As we can see, the average authority trust is the lowest among the four cases, as even when available there is a good probability that the reported information will not be correct. We have a low value of average social trust, but it is higher than the third case. Again, because of an unavailable and inaccurate authority, agents will rely on their self.

Let's then try to see the big picture, also comparing cases to each other.

In the first case (RS) we have the highest values of authority trust: it is a reliable available source, so that agents can rely on it. The authority trust has a good value also in the US case meaning that, in order to be trustworthy, it is important to be available to citizens, even if not always with correct information.

Considering the RW and UW cases, they seem to be very similar. Here in fact regardless of authority's reliability, trust on the authority is very low. Even the average social trust seems to be the same in the RW and UW cases. It reaches its maximal point in the RS case, being the other two sources quite always right, and its minimal point in the US, when the authority reaches all the agents, but it spreads incorrect information.

Summarizing, the US case seems to be good from the authority's point of view, but it seems to have a negative social impact.

Taking into account performances, as expected the best case is the RS one; having just trustworthy sources, agents' performances are very high. Again the RW and UW cases, in which the authority is unavailable, are quite the same (actually the UW cases' values are a little bit lower) meaning that if the authority is unavailable, it is no more important how much competent it is. The worst case is the US one. Here we have that all the agents' performances decreases to their lowest value.

Notice that the event3 is the one who suffers more. To understand this phenomenon it necessary to take into accounts the table 20. Lets' compare what happen in case of critical event and of light event. In case of critical event:

1. 1/3 of the population will estimate a 90% probability of critical event;
2. 1/3 of the population will estimate a 80% probability of critical event;
3. 1/3 of the population will estimate a 50% probability of critical event.

Conversely, in case of light event we have that:

1. 1/3 of the population will estimate a 100% probability of light event in the 75% of times, and a 10% probability of light event in the 25% of times;
2. 1/3 of the population will estimate a 100% probability of light event in the 50% of times, and a 20% probability of light event in the 50% of times;

3. 1/3 of the population will estimate a 50% probability of light event;

Then, even if it is more difficult for agents to detect a light event, when detected they will have a 100% certainty. On the contrary it is easier to them to identify a critical event, but not with high level of certainty. This means that, when there is the influence of another information sources that report wrong information, its influence on agents will be stronger in case of critical event rather than light event.

Finally, one could ask if it is better to have a reliable authority but not always available (RW) or an unreliable authority that has a strong presence (US). These results clearly state that the RW case is better, considering citizens' performance. This is due to the fact that, even if each individual citizen will receive right information from the authority about the 27% of the times in the RW case and about the 50% of the times in the US case, in the RW case the positive effect of the authority is widespread by the social effect. Then even if the authority is not doesn't reach everyone directly, it can count on the social effect to do it.

5. CONCLUSION

In this work we presented an articulated platform for social simulation, particularly suited for studying agents' choice in presence of critical weather phenomena. Using this framework, we were able to show some interesting results.

Through the training phase the agents learn to attribute to the different information sources the right values of trustworthiness and, as a consequence, they are able to perform quite effectively. In particular, two behaviors of the authorities are interesting: reliable and weakly communicative, not reliable and strongly communicative. They are a good simulation of the real cases in which the best prediction of a weather event is the more temporally close to the event itself (when becomes difficult to effectively spread the information: time for spreading is little). On the contrary, a prediction of a weather event can be effectively spread when there is big time for the spreading (far from the event), but this is in general a very inaccurate prediction.

Very interesting is the compensative and integrative role of the social phenomenon (observation of the others' behavior) that guides the performances of the agents upwards when just one of the two other sources results as reliable.

CHAPTER 6:

CONCLUSIONS

The main purpose of this research was to investigate how agents should handle information they possess in their decisional processes. This is a very easy task if they can access just one information source or they have many sources reporting the same information. But the problem becomes more complicated when we have to deal with multiple sources reporting different things. Who should the agent believe?

According with other authors (Melo et al, 2016) (Amgoud and Demolombe, 2014) (Villata et al, 2011) (Parsons et al, 2013) (Barber and Kim, 2001a) we state that using the concept of trust is an optimal way to face this problem. In fact it allows us to understand how much weight to give to each source. In a sense, trust on an information source allows us to subjectively evaluate the quality of information itself.

Then with this work we wanted to interpreter the role played by trust inside this decisional process.
A first phase of the work has been dedicated to a detailed analysis of all the possible cognitive variables that determine and influence this kind of trust. Then we developed a series of computational model, improving them in time until the final version that allows the presence of multiple information sources asserting different things.

This computational model has been studied by the means of a simulative tool in order to test its efficacy and to identify some useful practical outcomes of this theory. This allowed us to identify some interesting aspect both from the theoretical level and the practical level.

We started analyzing the concept of category applied to agents' evaluation. We wanted to understand how much this dimension is important for producing trust evaluations.

This work started from the consideration that he role of generalized knowledge has proven to determine the possibility to anticipate the value of unknown agents. Using just past experience I cannot evaluate a source I have never used. But if I know that the source belongs to a given category and I can produce/get/receive the evaluation of this category, then I can exploit this information to evaluate the source, in order to estimate how much it is trustworthy.

We tested the use of categories exploring some simulative scenario and then testing agents' evaluations. Inquiring these factors, the analysis provided in our simulations offers the possibility of better understanding and predicting the agent's behaviors.

Thanks to this analysis we proved that using both the past experience and the category it is possible to produce a better evaluation of an information source. Plus, this evaluation can be produced in even less time.

Another important result showed that, from the trustor point of view, it exists an optimal quantity of past experience to memorize. In fact memorizing just too few experiences does not allow modeling properly the source trustworthiness. However even memorizing too many past interactions has drawbacks. One risks in fact that the recorded information is too old, so that it does not reflect the current situation anymore. In this case the trustor will also need a lot of time to adjust its evaluation (and in that time the source performance could change again!).

We showed how to deduce a category evaluation from past experience when the first one is not available.
At last, in some cases it is possible that the categorization process has not been done properly (with a sufficient level of details) so that the category evaluations are too coarse, including members that are so much different from each other that the category itself loses its value. We showed how in this case it is still possible to successfully exploit categories, applying a clusterizzation process on the sources based on their performance. This is still valid if there are no predefined categories, so that it is possible to create new categories.

After a first analysis of the concept of category, we went beyond it, trying to apply it to recommendation. In the classical form of recommendation, an agent X provides a recommendation to an agent Y about an individual agent Z concerning how Z performs a given task. This definition can be extended introducing the concept of category. In particular an agent X can provide a recommendation to an agent Y about a whole category C of agents concerning how its members performs a given task on average.

According to this definition, we investigated the different roles that recommendations about individual agents and about categories of agents can play. Further, we analyzed the difference between this two kind of recommendation both in human society and in digital society. They are in fact completely different form a social point of view. The first and immediate kind of difference regards physical distance: in the human society communication happens to be limited by distance; this constrain does not stand anymore in the digital society, in which one agent can communicate with the entire world, without physically moving from its location.

The consequence of this two different kind of communication is that in the first one an agent/trustor will frequently communicate with the same agents (its neighbors), while in the second one the agent/trustor, as it is in the possibility to contact a lot of other agents, will communicate with more agents.

This kind of analysis can be particularly relevant to decide how to built the cognitive approach of agents searching information among multiple sources. Before choosing between direct or generalized information, we need to evaluate how information is distributed among the agents in the specific domain. Our results show that in certain cases becomes essential the use of categorial knowledge for selecting qualified partners. In particular when the quantity of information (about the agents' trustworthiness exchanged in the system) decreases it is better to rely on the categorial recommendations rather than individual recommendations.

Concerning the human and digital society, we discovered that the utility of categories is higher in the digital world case.

In fact, in the digital society the agents can have access to more other agents, as they are not constrained by physical distance. In this way, they know more agents, but their knowledge about each single agent is limited.

Conversely in the human society, each agent can ask just to its neighbors. Although they move into the world, their knowledge is strictly related to their physical position. As a consequence, they will know better their neighbors and their knowledge of categories strongly depends on the individuals they have met.

Our last work focuses on how cognitive agents decide in presence of different hydrogeological phenomena, analyzing also how they use their information source and learn how much trustworthy they are.
A result of this work is the realized platform itself. In fact setting its parameters, it can be manipulated allowing shaping many possible scenarios. Using this platform, we were able to show some interesting result.

The simulations also showed how, on the basis of a training phase in these different situations, the agents were able to make a rational use of their different information sources. They learnt to attribute to the different information sources the right values of trustworthiness and, as a consequence, they are able to perform quite effectively.
Focusing on the authority, it is almost obvious that the best case is the one in which it can reach every one with a high degree of reliability. However this case is quite unrealistic. In a concrete case weather forecast's accuracy is indissolubly linked by time constrains: they improve while the event is approaching. Clearly it is necessary a trade-off between reliability and communicativeness.

In particular, two behaviors of the authorities are interesting: reliable and weakly communicative, not reliable and strongly communicative. They are a good simulation of the real cases in which the best prediction of a weather event is the more temporally close to the event itself (when becomes difficult to effectively spread the information: time for spreading is little). On the contrary, a prediction of a weather event can be effectively spread when there is big time for the spreading (far from the event), but this is in general a very inaccurate prediction.

Another interesting effect is produced by the social source, which derives from the observation of the others' behavior. It has a compensative and integrative role that guides the performances of the agents upwards when just one of the two other sources results as reliable.

Resuming, from this work underlines the importance of managing information sources. The proposed approach makes use of the concept of trust to handle them. This is reasonable as the literature shows.

While applying this model, it emerged the utility of using categories for trust evaluation, as it is and applied to recommendations. Even if the there is a lot of literature about categories, the idea of category recommendation is still novel and it should be studied more. As proved, it could have a very strong impact in the emerging web society.

We also showed how this theory can be applied in a practical context. We chose the one of hydrogeological phenomena, as the presence of critical situation is particularly interesting for studying the decisional dynamics. Even here there would be a lot to say. First of all, it would be interesting to analyze the influence of the authority's behavior on the population: which is the best behavior? How can it maximize citizens' performance? Is it better to have a scaremonger or a prudent authority?

It is also interesting to study the social dynamics that exists among citizens: how are they influence by others? How do they influence others? Have they different levels of sociality? Can this social source compensate the lack of other information? How does impulsivity influence their choice?

One could also take into account the relationships among citizens. For instance (De Meo et al, 2015) study groups exploiting the concept of compactness, which they define as a combination of similarity and trust among agents.

All of them are interesting open questions that could be investigated in the future.

REFERENCES

1. Adomavicius, G., Tuzhilin, A. (2005) Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering (TKDE) 17, 734–749.
2. L. Amgoud, R. Demolombe, An Argumentation-based Approach for Reasoning about Trust in Information Sources , In *Journal of Argumentation and Computation, 5(2), 2014*
3. (Barber and Kim, 2001a) Barber, K. S., & Kim, J. (2001). Belief revision process based on trust: Agents evaluating reputation of information sources. In *Trust in Cyber-societies* (pp. 73-82). Springer Berlin Heidelberg.
4. (Barber and Kim, 2001b) Barber, K. S., & Kim, J. (2001). Belief revision process based on trust: Simulation experiments. In *In Proceedings of Autonomous Agents' 01 Workshop on Deception, Fraud, and Trust in Agent Societies*.
5. Burnett, C., Norman, T., and Sycara, K. 2010. Bootstrapping trust evaluations through stereotypes. In Pro- ceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'10). 241248.
6. Burnett, C., Norman, T., and Sycara, K. (2013) Stereotypical trust and bias in dynamic multiagent systems. ACM Transactions on Intelligent Systems and Technology (TIST), 4(2):26, 2013.
7. Capra, C. M. (2004). Mood-driven behavior in strategic interactions. *The American Economic Review*, *94*(2), 367-372.
8. Castelfranchi, C., and Falcone, R. 2000. Trust is much more than subjective probability: Mental components and sources of trust. In

Proceedings of the 33rd Hawaii International Conference on System Science. Maui, Hawai'i: IEEE Com- puter Society.

9. Castelfranchi, C., Falcone R., Pezzulo, (2003) Trust in Information Sources as a Source for Trust: A Fuzzy Approach, Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-03) Melburne (Australia), 14-18 July, ACM Press, pp.89-96.
10. Castelfranchi C., Falcone R., Trust Theory: A Socio-Cognitive and Computational Model, John Wiley and Sons, April 2010.
11. Castelfranchi C., Falcone R., Sapienza A., Information sources: Trust and meta-trust dimensions. In proceeding of the workshop TRUST 2014, colocated with AAMAS 2014, Paris, CEUR Workshop Proceedings 2014 (In press)
12. Conte R., and Paolucci M., 2002, Reputation in artificial societies. Social beliefs for social order. Boston: Kluwer Academic Publishers.
13. P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Improving Recommendation Quality by Merging Collaborative Filtering and Social Relationships. In Proc. of the International Conference on Intelligent Systems Design and Applications (ISDA 2011) , Córdoba, Spain, IEEE Computer Society Press, 2011
14. De Meo, P., Ferrara, E., Rosaci, D., & Sarné, G. M. (2015). Trust and compactness in social network groups. *IEEE transactions on cybernetics*, *45*(2), 205-216.
15. Demolombe R., (1999), To trust information sources: A proposal for a modal logic framework. In Castelfranchi C., Tan Y.H. (Eds), Trust and Deception in Virtual Societies. Kluwer, Dordrecht.
16. Demolombe, R. (2004, March). Reasoning about trust: A formal logical framework. In *International Conference on Trust Management* (pp. 291-303). Springer Berlin Heidelberg.

17. Falcone R., Castelfranchi, C. (2004), Trust Dynamics: How Trust is influ- enced by direct experiences and by Trust itself; Proceedings of the 3rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-04), New York, 19-23 July 2004, ACM-ISBN 1-58113-864-4, pages 740-747.
18. Falcone R, Castelfranchi C, (2008) Generalizing Trust: Inferencing Trustworthiness from Categories. In Proceedings, pp. 65 - 80. R. Falcone, S. K. Barber, J. Sabater-Mir, M. P. Singh (eds.). Lecture Notes in Artificial Intelligence, vol. 5396. Springer, 2008
19. Falcone R., Castelfranchi C., Trust and Transitivity: How trust-transfer works, 10th International Conference on Practical Applications of Agents and Multi-Agent Systems, University of Salamanca (Spain)28-30th March, 2012.
20. Falcone R., Piunti, M., Venanzi, M., Castelfranchi C., (2013), From Manifesta to Krypta: The Relevance of Categories for Trusting Others, in R. Falcone and M. Singh (Eds.) Trust in Multiagent Systems, ACM Transaction on Intelligent Systems and Technology, Volume 4 Issue 2, March 2013
21. Fang H., Zhang J., Sensoy M., and Thalmann N. M. (2012) A generalized stereotypical trust model. In Proceedings of the 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pages 698–705, 2012.
22. Gambetta, Diego (2000) ‘Can We Trust Trust?’, in Gambetta, Diego (ed.) Trust: Making and Breaking Cooperative Relations, electronic edition, Department of Sociology, University of Oxford, chapter 13, pp. 213-237, http://www.sociology.ox.ac.uk/papers/gambetta213-237.pdf
23. Guo, G., Zhang, J., & Yorke-Smith, N. (2015). Leveraging multiviews of trust and similarity to enhance clustering-based recommender

systems. *Knowledge-Based Systems*, *74*, 14-27.

24. Hertzum, M., Andersen, H. H., Andersen, V., & Hansen, C. B. (2002). Trust in information sources: seeking information from people, documents, and virtual agents. *Interacting with computers*, *14*(5), 575-599.
25. Higgins, E. T. (1997). Beyond pleasure and pain. American Psychologist, 52, 1280-1300.
26. Huynh, T.D., Jennings, N. R. and Shadbolt, N.R. , 2006, An integrated trust and reputation model for open multi-agent systems. Journal of Autonomous Agents and Multi-Agent Systems, 13, (2), 119-154.,
27. S. Jiang, J. Zhang, and Y.S. Ong. An evolutionary model for constructing robust trust net- works. In Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2013.
28. Kahneman, Daniel, and Amos Tversky, ”Prospect Theory: An Analysis of Decision Under Risk”. Econometrica. XLVII (1979): 263-291.
29. King, D.W., Casto, J., Jones, H., 1994. Communication by Engineers: A Literature Review of Engineers' Information Needs, Seeking Processes, and Use, Council on Library Resources, Washington, DC.
30. B. Liu, Uncertainty theory 5th Edition, Springer 2014.
31. Lops P., Gemmis M., and Semeraro G., (2011), “Content-based recommender systems: State of the art and trends,” in Recommender Systems Handbook. Springer, pp. 73–105.
32. P. Massa, P. Avesani, Trust-aware recommender systems, RecSys '07: Proceedings of the 2007 ACM conference on Recommender systems, 2007
33. Matt, P. A., Morge, M., & Toni, F. (2010, May). Combining statistics and arguments to compute trust. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent*

Systems: volume 1-Volume 1 (pp. 209-216). International Foundation for Autonomous Agents and Multiagent Systems.

34. Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.
35. Melaye, D., & Demazeau, Y. (2005). Bayesian dynamic trust model. In Multi-agent systems and applications IV (pp. 480-489). Springer Berlin Heidelberg.
36. Melo, Victor S., Alison R. Panisson, and Rafael H. Bordini. "Trust on Beliefs: Source, Time and Expertise.", in *Proceedings of the 18th International Workshop on Trust in Agent Societies co-located with the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), Singapore, May 10, 2016,* Ceur Workshop Proceedings, vol 1578, paper 6.
37. Messina, F., Pappalardo, G., Rosaci, D., Santoro, C., & Sarné, G. M. L. (2016). A trust-aware, self-organizing system for large-scale federations of utility computing infrastructures. *Future Generation Computer Systems*, *56*, 77-94.
38. Mutti, A. (1987). La fiducia. Un concetto fragile, una solida realtà. *Rassegna italiana di sociologia*, *28*(2), 223-247.
39. Myers, D., & Tingley, D. (2011). The influence of emotion on trust. *Political Analysis, Forthcoming*.
40. Paglieri F., & Castelfranchi C. (2012). Trust in relevance. In: S. Ossowski, F. Toni, G. Vouros (Eds.), Proceedings of the First International Conference on Agreement Technologies (AT 2012). CEUR Workshop Proceedings, vol. 918: CEUR-WS.org, pp. 332 - 346.
41. Parsons, S., Sklar, E., Singh, M. P., Levitt, K. N., & Rowe, J. (2013, March). An Argumentation-Based Approach to Handling Trust in

Distributed Decision Making. In *AAAI Spring Symposium: Trust and Autonomous Systems*.

42. Pinelli, T.E., Bishop, A.P., Barclay, R.O., Kennedy, J.M., 1993. The information-seeking behavior of engineers. In: Kent, A., Hall, C.M. (Eds.), Encyclopedia of Library and Information Science, vol. 52. Marcel Dekker, New York, pp. 167–201.
43. Quercia, D., Hailes, S., & Capra, L. (2006). B-trust: Bayesian trust framework for pervasive computing. In Trust management (pp. 298-312). Springer Berlin Heidelberg.
44. Ramchurn S., Jennings N., Sierra C., and Godo L. (2004) Devising a trust model for multi-agent interactions using confidence and reputation. Applied Artificial Intelligence, 18(9-10):833-852.
45. Sabater-Mir J., Sierra C., (2001), Regret: a reputation model for gregarious societies. In 4th Workshop on Deception and Fraud in Agent Societies (pp. 61-70). Montreal, Canada.
46. Sabater-Mir, J. 2003. Trust and reputation for agent societies. Ph.D. thesis, Universitat Autonoma de Barcelona.
47. Sapienza, A., Falcone, R., & Castelfranchi, C. Trust on Information Sources: A theoretical and computation approach, in proceedings of WOA 2014, ceur-ws, vol 1260, paper 12.
48. Sensoy M., Yilmaz B., and Norman T. J. 2014, STAGE: Stereotypical trust assessment through graph extraction. Computational Intelligence.
49. Sharpanskykh, A., & Treur, J. (2010, August). Behavioural abstraction of agent models addressing mutual interaction of cognitive and affective processes. In *International Conference on Brain Informatics* (pp. 67-77). Springer Berlin Heidelberg.
50. Sowell, T. (1980). *Knowledge and decisions*. Basic Books.

51. Sztompka, P. (1999). *Trust: A sociological theory*. Cambridge University Press.

52. C. Than and S. Han, Improving Recommender Systems by Incorporating Similarity, Trust and Reputation, Journal of Internet Services and Information Security (JISIS), volume: 4, number: 1, pp. 64- 76, 2014

53. Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira, Computing Confidence Values: Does Trust Dynamics Matter? In L. Sabra Lopes et al. (Eds.): EPIA 2009, LNAI 5816, pp. 520-531, 2009, Springer.

54. Villata, S., Boella, G., Gabbay, D. M., & Van Der Torre, L. (2011, June). Arguing about the trustworthiness of the information sources. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty* (pp. 74-85). Springer Berlin Heidelberg.

55. D. Walton, Argumentation Schemes for Presumptive Reasoning , Mahwah, N.J., Lawrence Erlbaum Associates, 1996.

56. D. Walton, C. Reed and F. Macagno, Argumentation Schemes, Cambridge, Cambridge University Press, 2008.

57. Wang, Y., & Vassileva, J. (2003, October). Bayesian network-based trust model. In Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on (pp. 372-378). IEEE.

58. Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

59. Yolum, P. and Singh, M. P. 2003. Emergent properties of referral systems. In Proceedings of the 2nd International Joint Conference on Autonomous Agents and MultiAgent Systems (AAMAS'03).