Trust in Information Sources as a Source for Trust: A Fuzzy Approach[1]

Cristiano Castelfranchi
Istituto di Scienze e Tecnologie della Cognizione - CNR
viale Marx, 15 - Roma
++39 06 86890518
castel@ip.rm.cnr.it

Rino Falcone
Istituto di Scienze e Tecnologie della Cognizione - CNR
viale Marx, 15 - Roma
++39 06 86890211
falcone@ip.rm.cnr.it

Giovanni Pezzulo
Istituto di Scienze e Tecnologie della Cognizione - CNR
viale Marx, 15 - Roma
++39 06 86890208
pezzulo@ip.rm.cnr.it

ABSTRACT

The aim of this paper is to show how relevant is a trust model based on beliefs and their credibility.

The approaches to the study of trust are various and very different from each of other. In our view, just a socio-cognitive approach to trust would be able to analyse the sub-components on which the final decision to trust or not is taken. In this paper we show an implementation of our socio-cognitive model of trust developed using the so-called Fuzzy Cognitive Maps. The model allows to distinguish between internal and external attributions and it introduced a degree of trust derived from the credibility of the trust beliefs, while the credibility of the beliefs derives from their sources and the sources' number, convergence, reliability (i.e. trust).

With this implementation we show how the different components may change and how their impact can change depending on the specific situation and from the agent heuristics or personality. In particular, we analyse the different nature of the belief sources and their trustworthiness. We assumed different types of belief sources. For each trustier's belief one should consider what the content of the belief is, who/what the source is, how this source evaluates the belief, how the trustier evaluates this source (with respect to this belief). In addition for considering the contribution of different sources we need a theory of how they combine. The interesting thing in this paper is that starting from finding the sources of trust we are obliged to consider the trustworthiness of these sources**.**

General Terms: Experimentation, Theory

Keywords: Trust, Beliefs, Sources of beliefs, Fuzzy Cognitive Maps, Medical House Assistance Scenarios.

AAMAS'03, July 14–18, 2003, Melbourne, Australia.

1. INTRODUCTION

In this paper we will show a possible implementation and advance of the socio-cognitive model of trust developed in [1, 2]. This implementation uses a fuzzy approach (in particular, it uses the so-called Fuzzy Cognitive Maps - FCM [3]).

The aim of this paper is to show how relevant is a trust model based on beliefs and their credibility. In addition, given that the credibility of a belief directly depends from the credibility of its sources, we also analyse the different nature of the belief sources and their trustworthiness.

The richness of the referred model (trust is based on many different beliefs) allows to distinguish between internal and external attributions (to the trustee) and for each of these two attributions it allows to distinguish among several other sub-components such as: competence, disposition, unharmfulness and so on. In fact, our model introduced a degree of trust instead of a simple probability factor since it permits to evaluate the trustfulness in a rational way. In other words, trust can be said to consist of or better to (either implicitly or explicitly) imply the *subjective probability* (in the sense of a subjective evaluation and perception of the risks and opportunities) of the successful performance of a given behavior, and it is on the basis of this subjective perception/evaluation that the agent decides to rely or not, to bet or not on the trustee. However, the probability index is based on, derives from those beliefs and evaluations. In other terms the global, final probability of the realisation of the goal *g*, i.e. of the successful performance of an action α, should be decomposed into the probability of the trustee performing the action well (that derives from the probability of willingness, persistence, engagement, competence: *internal attribution*) and the probability of having the appropriate conditions (opportunities and resources *external attribution*) for the performance and for its success, and of not having interferences and adversities (*external attribution*). Why this decomposition is important? Not only for cognitively grounding such a probability - and this cognitive embedding is fundamental for relying, influencing, persuading, etc.-, but because:

a) the agent's trusting decision might be different with the same global probability or risk, depending on its composition;

[1] This paper has been partially founded by the ALFEBIITE European Project (IST-1999-10298); by the Progetto MIUR 40% "Agenti software e commercio elettronico" and by the PAR project of the University of Siena.

b) trust composition (internal vs external) produces completely different intervention strategies: manipulating the external variables (circumstances, infrastructures) is completely different than manipulating internal parameters.

In such a way we understand how the attribution of trust is a very complex task, and that the decision making among different alternative scenarios is based on a complex evaluation of the basic beliefs and of their own relationships. And again, how the (even minimal) change of the credibility value of any (very relevant) belief might influence the resulting decision (and thus the trustworthiness attributed to the trustee); or vice versa, how significant changes in the credibility value of any unimportant belief does not modify the final trust.

2. WHY THE FUZZY APPROACH

We have chosen an approach based on the Fuzzy Logic for several reasons. First, we want to model some graded phenomenon like trust that is difficult to estimate experimentally. The qualitative approach of the Fuzzy Logic is very useful because it is intuitive to start the analysis with natural language labels (this doctor is *very skilled*) that represent intervals rather than exact values. More, the behavior of these systems (e.g. their combinatorial properties) seems to be good in modelling several cognitive dynamics [3], even if to find "the real function" for a mental operation and to estimate the contribution of convergent and divergent belief sources remain open problems.

We have used an implementation based on a special kind of fuzzy system called Fuzzy Cognitive Maps (FCM); they allow to compute the value of the trustfulness starting from belief sources that refer to trust features. The values of those features are also computed, allowing us to perform some cognitive operations that lead to the effective decision to trust or not to trust (e.g. impose an additional threshold on a factor, for example risks). Using this approach we describe beliefs and trust features as approximate (mental) objects with a strength and a causal power one over another.

3. SCENARIOS

In order to exemplify our approach and system we will apply it to an interesting scenario, that is one of the application scenarios identified within the Alfebiite Project [4]. The scenario we are going to study is medical house assistance in two particular instances: a) a doctor (a human operator) visiting a patient at home and b) a medical automatic system for supporting the patient (without direct human intervention).

The case studies under analysis are:

- an **emergency situation**, in which there is the necessity of identifying an occurring danger (for example, a hearth attack) as soon as possible to cope with it; we consider in this case the fact that the (first) therapy to be applied is quite simple (suppose just a injection);

- a **routine situation**, in which there is a systematic and specialist therapy to apply (with quite a complex procedure) but in which there is no immediate danger to cope with.

We will show how the factors that produce the final trust for each possible trustee are dependent on:

- the initial strength of the different beliefs (on which trust is based) but also

- how much a specific belief impacts on the final trust (the causality power of a belief).

It is through this second kind of factors that we have the possibility also of characterizing some personality traits of the agents [5, 6, 7].

4. BELIEF SOURCES

In our model trust is an "*evaluation*" and an "*expectation*" (i.e. in our theory special kinds of beliefs) and also an (affective) attitude and disposition. They are based upon more specific beliefs which are both *basis* of trust and its *sub-components* or *parts*: which/how is our trust in (evaluation of) the trustee as for his/her/its competence and ability? Which/how is our trust in (evaluation of) the trustee as for his/her/its intention and reliability? Which/how is our trust in (evaluation of) the trustee as for his/her/its goodwill and honesty? And so on.

Those beliefs are the analytical account and the components of trust, and we derive *the degree of trust* directly from the *strength* of its componential and supporting beliefs. More precisely in our model [2] we claim that *the degree of trust is a function of the subjective certainty of the pertinent beliefs*. We used the degree of trust to formalize a rational basis for the decision of relying and betting on the trustee. Also in this case we claimed that the "quantitative" aspect of another basic ingredient is relevant: *the value or importance or utility of the goal g*, will obviously enter the evaluation of the risk, and will also modify the required threshold for trusting. In sum, *the quantitative dimensions of trust are based on the quantitative dimensions of its cognitive constituents*.

It should be clear that in our view trust is not an arbitrary index just with an operational importance, without a real content, but it is based on the subjective certainty of the pertinent beliefs. However, what is the origin and the justification of the strength of beliefs? Just their sources. In our theory, depending on the nature, the number, the convergence/divergence, and the credibility of its sources a given belief is more or less strong (certain, credible).

Several models propose a quantification of the degree of trust and make it dynamic, i.e. they can change and update such a degree [8, 9]. But they only consider direct interaction (experience) or reputation as sources. In this paper we have considered four possible types of belief sources: *direct experience* (how the personal –positive or negative- experience of the trustier contributes to that belief); *categorization* (how the properties of a class are transferred to their members); *reasoning* (more general than just categorization); and *reputation* (how the other's experience and opinion influences the trustier beliefs). The dynamic of this model does not consider the possibility of learning. We are just modelling the resulting effects that a set of trustier's basic beliefs (based on various sources) have on the final trustfulness of the trustee about a given task and in a specific situation. At present we do not consider how these effects feedback on the basic beliefs.

4.1 BUILDING BELIEF SOURCES

Agents act depending on what they believe, i.e. *relying on* their beliefs. And they act on the basis of the degree of reliability and certainty they attribute to their beliefs. In other words, trust/confidence in an action or plan (reasons to choose it and expectations of success) is grounded on and derives from trust/confidence in the related beliefs. We have assumed four

types of belief sources: Direct Experience, Categorization, Reasoning, and Reputation. For each of these kinds of sources we have to consider the impact it produces on trustier's beliefs about trustee's features. These impacts result from the composition of the value of the content (property) of that specific belief (the object belief) with a subjective modulation introduced by some epistemic evaluations about that specific source. In fact when we have a belief we have to evaluate:

- the value of the content of that belief;

- who/what the source is (another agent, my own inference process, a perceptive sense of mine, etc.);

- how this source evaluates the belief (the subjective certainty of the source itself);

- how the trustier evaluates this source (with respect to this belief).

Those beliefs are not all at the same layer. Clearly some of them are meta-beliefs, and some of them tune, modulate the value and the impact of the lower beliefs. The general schema could be described as a cascade having two levels (see Figure1); at the bottom level there is the single belief (in particular, the value of the content of that specific belief; this value should be used (have a part) in the trustier's evaluation of some trustee's feature); at the top level there is the composition of the previous value with the epistemic evaluations of the trustier. At this level all the contributions of various sources of the same type are integrated.

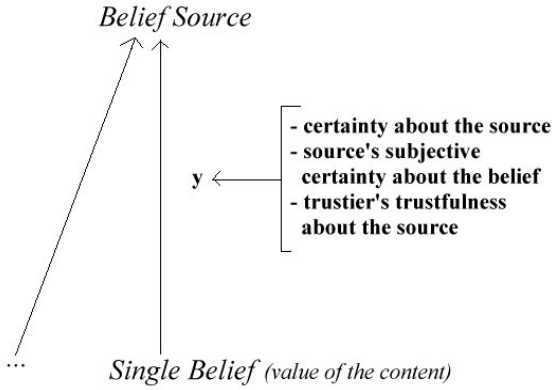

Figure 1. From single beliefs to the belief source

Let us consider as an example the belief source of Reputation about a doctor's Ability (see Figure 2). In order to have a value, we have to consider many opinions about the ability of that doctor. For example John may have an opinion: *I think that the doctor is quite good at his work*. In this case we have the belief "*the doctor is quite good at his work*" and the belief source: "*John*". Considering, in this specific case, the four factors above described, we have: the value of the content (doctor is *quite good*); how good John considers his own belief ("I think" that could mean: *I am sure/ I am quite sure/ I am not so sure* and so on); the degree of certainty that John has expressed this opinion (*I am sure* that John told me (thinks) that, etc.); the credibility of John's opinion.

The first factor represent a property, a belief and the value of its content (for example ability); it is a source's belief that becomes an object of the trustier's mental world. The second factor represents an epistemic evaluation that the source does on the communicated belief. The third factor represents trustier's degree of certainty that the source expressed (communicated) that belief (it is also linked with the trustier's selftrust). Finally, the fourth factor represents a degree of trust in source's opinion, and it depends on a set of trustier's beliefs about source's credibility, ability to judge and so on.

The second, third and the fourth factors are not objects of the same level, but rather meta-beliefs: they represent a *modulation* of the beliefs. In our networks, this can be better represented as impact factors. So, in our network we have two main nodes: "John's belief" and "Reputation about ability". The first factor sets the value of the first node. The second, third and fourth factors set the value of the edge from the first to the second node.

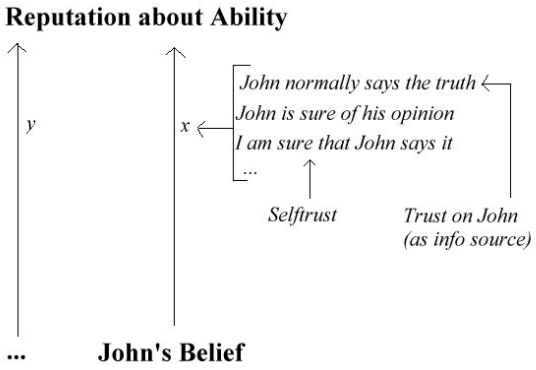

Figure 2. Case of belief source of Reputation.

4.2 CONVERGING AND DIVERGING BELIEF SOURCES

In order to consider the contribution of different sources we need a theory of how they combine. The combination of different information sources is a classical complex problem [10, 11]. It is in particular an evident problem in the case in which we are going to model human behaviours. In fact, humans use very different strategies and mechanisms, essentially based on their personalities and experiences. This problem is very relevant in the case in which there are diverging opinions (beliefs). In these cases humans could use various heuristics for combining the opposite values even because the elements which should be combined could produce an incoherent picture: if someone says that Mary dresses a hat and another one says that she does not dress a hat, I cannot infer that Mary dresses half a hat; or again if there are two persons that both say that Doctor Smith is a not too good and not too bad doctor while other two persons give us two diverging evaluations on Doctor White (one says that he is an excellent doctor and another says that he is a really bad doctor) we would not have an equivalent evaluation of Doctor White and Doctor Smith, and our decision would be guided by other criteria. These criteria are linked with context, emotions, personality factors. We could have people that in the presence of diverging opinions decide to suspend the judgment (they become unable to decide), or people that take in consideration the best opinion (optimistic personality), or, on the contrary, people that take in consideration the worst opinion (pessimistic personality). And so on. A good model should be able to implement different heuristics. For the moment, in our model we simply sum up all the contributions and we squash the result with a threshold function. In fact, the exact heuristics that humans choose depend on the situation and

eventually the exact threshold functions can be object of empirical analysis or simulations. The model itself is independent from those heuristics, that can be easily substituted.

4.3 HOMOGENOUS AND HETEROGENEOUS SOURCES

We have the problem of summing up the contribution of many different sources. We have already discussed the case of homogenous sources (e.g. different opinions about a feature/person/thing, etc.), when an heuristic has to be chosen. The same problem occurs when we want to sum up the contribution of heterogeneous fonts (e.g. direct experience and reputation about the ability of a doctor). Even in this case, many heuristics are possible. For example, which is it more relevant our own personal experience or the reputation about a specific ability of a person? There is not a definitely right answer to this question: are we able to evaluate that ability in a good way? Or is it better to rely on the evaluation of others? And vice-versa. Our analysis is limited to a plain estimation of all the relevant factors, but many other strategies are possible, as in the case of homogenous sources. Also in this case, some strategies depend on personality factors.

We have described how it is possible to model belief sources starting from the single beliefs; now we describe how trust is computed starting from the belief sources.

5. MODELING BELIEFS AND SOURCES

Following a belief-based model of trust [1] we can distinguish between trust in the trustee (be it either someone –e.g. the doctor- or something –e.g. the medical automatic system-) that has to act and produce a given performance thanks to its internal characteristics, and the (positive and/or negative) environmental conditions (like opportunities and interferences) affecting the trustee's performance, that we call "external factors".

In this paper we take into account:

- Three main beliefs regarding the trustee: an ability/competence belief; a disposition/availability belief, and an unharmfulness belief.

- Two main beliefs regarding the contextual factors: opportunity beliefs and danger beliefs.

5.1 BELIEFS AND SOURCES OVERVIEW

Which are the meanings of our basic beliefs in the case of the doctor and in the case of the medical automatic system?

For the *medical automatic system* the internal and external factors that we consider are:

- *Internal factors – ability/competence beliefs*: these beliefs concern the efficacy and efficiency of the machine; its capability to successfully apply the right procedure in the case of correct/proper use of it. Possibly, also its ability to recover from an inappropriate use.

- *Internal factors – disposition/availability beliefs*: these beliefs are linked to the reliability of the machine, its regular functioning, its easiness of use; possibly, its adaptability to new and unpredictable uses.

- *Internal factors – unharmfulness beliefs*: these beliefs concern the absence (lack) of the internal/ intrinsic risks of the machine: the dangers implied in the use of that machine (for example side effects for the trustier's health), the possibility of breaking and so on.

- *External factors – opportunity beliefs*: concerning the opportunity of using the machine, independent of the machine itself, from the basic condition to have the room for allocating the machine to the possibility of optimal external conditions in using it (regularity of electric power, availability of an expert person in the house that might support in its use, and so on).

- *External factors – danger beliefs*: these beliefs are connected with the absence (lack) of the systemic risks and dangers external to the machine that could harm the user: consider for example the risk for the trustier's privacy: in fact we are supposing that the machine is networked in an information net and the data are also available to other people in the medical structure.

For the *doctor* the internal and external factors that we consider are:

- *Internal factors – ability/competence beliefs*: these beliefs concern the (physical and mental) skills of the doctor; his/her ability to make a diagnosis and to solve problems.

- *Internal factors – disposition/availability beliefs*: these beliefs concern both the willingness of the doctor to commit to that specific task (subjective of the specific person or objective of the category), and also his/her availability (in the sense of the possibility to be reached/informed about his/her intervention).

- *Internal factors – unharmfulness beliefs*: these beliefs concern the absence (lack) of the risks of being treated by a doctor; namely the dangers of a wrong diagnosis or intervention (for example, for the health of the trustier).

- *External factors – opportunity beliefs*: concerning the opportunities not depending on the doctor but on conditions external to his/her intervention. Consider for example the case in which the trustier is very close to a hospital in which there is an efficient service of fast intervention; or again, even if the trustier is not very close to a hospital he/she knows about new health policies for increasing the number of doctors for quick intervention; and so on. Conversely, imagine a health service not efficient, unable to provide a doctor in a short time; or, again, a particularly chaotic town (for the car traffic, for the frequent strikes in it) that could hamper the mobility of the doctors and of their immediate transfer in the site where the patient is.

- *External factors – danger beliefs*: these beliefs concern with the absence (lack) of the risks and dangers which do not depend directly on the doctor but on the conditions for his/her intervention: for instance, supposing that the trustier's house is poor and not too clean, the trustier could see the visit of a person (the doctor in this case) as a risk for his/her reputation.

Each of the above mentioned beliefs may be generated by different sources; such as: direct experience, categorization, reasoning, and reputation. So, for example, ability/competence beliefs about the doctor, may be generated by the direct knowledge of a specific doctor, and/or by the generalized knowledge about the class of doctors and so on.

6. OVERVIEW OF THE IMPLEMENTATION

We describe an implementation that uses Fuzzy Cognitive Maps (FCM) [3]. An FCM is an additive fuzzy system with feedback; it is well suited for representing a dynamic system with cause-effect

relations. An FCM has several nodes, representing causal concepts (belief sources, trust features and so on), and edges, representing the causal power of a node over another one. The values of the nodes representing the belief sources and the values of all the edges are assigned by a human; these values propagate in the FCM until a stable state is reached; so the values of the other nodes (in particular the value of the node named Trustfulness) are computed. In order to design the FCM and to assign a value to its nodes we need to answer four questions: which value do I assign to this concept? How sure am I of my assignment? Which are the reasons of my assignment? How much does this concept impacts on an other linked concept?

We address the first and the second question above assigning numeric values to the nodes representing the belief sources. The nodes are causal concepts; their value varies from –1 (true negative) to +1 (true positive). This number represents the value/degree of each single trust feature (say ability) by combining together both the credibility value of a belief (degree of credibility) and the estimated level of that feature. Initial values are set using adjectives from natural language; for example, "I believe that the ability of this doctor is *quite good* (in his work)" can be represented using a node labeled "ability" with a little positive value (e.g. +0.4). For example, the value +0.4 of ability either means that the trustier is *pretty sure* that the trustee is *rather good*, or that he/she is *rather sure* that the trustee is *really excellent*, etc.

We address the third question above designing the graph. Some nodes receive input values from other nodes; these links represent the reasons on which their values are grounded. Direct edges stand for fuzzy rules or the partial causal flow between the concepts. The sign (+ or -) of an edge stands for causal increase or decrease. For example, the Ability value of a doctor influences positively (e.g. with weight +0.6) his Trustfulness: if ability has a positive value, Trustfulness increases; otherwise it decreases.

We address the fourth question above assigning values to the edges: they represent the impact that a concept has over another concept. The various features of the trustee, the various components of trust evolution do not have the same impact, and importance. Perhaps, for a specific trustee in a specific context, ability is more important than disposition. We represent the different quantitative contributions to the global value of trust through these weights on the edges. The possibility of introducing different impacts for different beliefs surely represents an improvement with respect to the trust basic model. FCMs allow to quantify causal inference in a simple way; they model both the strength of the concepts and their relevance for the overall analysis. For example, the statement: "Doctors are *not very accessible* and this is an *important factor* (for determining their trustfulness) in an emergency situation" is easily modelled as a (strong) positive causal inference between the two concepts of Accessibility and Trustfulness. FCMs also allow to sum up the influence of different causal relations. For example, adding another statement: "Doctors are *very good* as for their ability, but this is a *minor factor* in an emergency situation" means adding a new input about the Ability, with a (weak) positive causal influence over Trustfulness. Both Accessibility and Ability, each with its strength and its causal power, contribute to establish the value of Trustfulness.

6.1 A Note on Fuzzy Values

Normally in fuzzy logic some labels (mainly adjectives) from natural language are used for assigning values; each label represents a range of possible values. There is not a single universal translation between adjectives and the exact numerical values in the range. Differently from standard Fuzzy techniques, in FCM it is required to use crisp input values; we have used the average of the usual ranges, obtaining the following a of labels, both for positive and negative values: quite; middle; good; etc. However, as our experiments show, even with little variation of these values into the same range, the FCMs are stable and give similar results. As Figure 3 shows, the ranges we have used do not divide the whole range {-1,1} into equal intervals; in particular, near the center (value zero) the ranges are larger, while near the two extremities they are smaller. This implies that a little change of a value near the center normally does not lead to a "range jump" (e.g. from some to quite), while the same little change near the extremities can (e.g. from very to really). This topology is modeled in the FCM choosing the threshold function; in fact, it is possible to choose different functions, the only constraint is that this choice must be coherent with the final convergence of the algorithm. With the function chosen in our implementation, changes in big (positive or negative) values have more impact in the FCM, this is a tolerable result even if it is does not correspond with a general cognitive model.

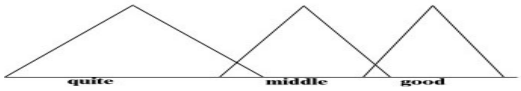

Figure 3. Fuzzy Intervals

7. DESCRIPTION OF THE MODEL

Even if FCMs are graphs, ours can be seen as having four layers. The first layer models the influence of the "beliefs sources" (as we have seen in §4): Direct Experience (e.g. "In my experience…"), Categorization (e.g. "Usually doctors…"), Reasoning (e.g. "I can infer that…"), Reputation (e.g. "A friend says that…"). Their value is meant to be stable (i.e. it does not change during computation), because these nodes could be assumed being the result of an "inner FCM" where each single belief is represented (e.g. Direct Experience about Ability results from many nodes like: "I was visited *many times* by this doctor and he was *really good* at his work", "*Once* he made a *wrong* diagnosis", …). So their value not only represents the strength of the feature expressed in the related beliefs, but also their number and their perceived importance, because belief sources represent the synthesis of many beliefs. The second layer shows the five relevant basic beliefs: Ability, Accessibility, Harmfulness, Opportunities and Danger. These basic beliefs are distinguished in the third layer into Internal Factors and External Factors. Ability, Accessibility and Harmfulness are classified as Internal Factors; Opportunities and Danger are classified as External Factors. Internal and External factors both influence Trustfulness, which is the only node in the fourth layer. For the sake of simplicity no crossing-layer edges are used, but this could be easily done since FCM can compute cycles and feedback.

7.1 Running the Model

Once the initial values for the first layer (i.e. belief sources) are set, the FCM starts running[2]. The state of a node N at each step s is computed taking the sum of all the inputs, i.e., the current values at step s-1 of nodes with edges coming into N multiplied by the corresponding edge weights. The value is then squashed (into the –1,1 interval) using a threshold function. The FCM run ends when an equilibrium is reached, i.e., when the state of all nodes at step s is the same as that at step s-1. At this point we have a resulting value for Trustfulness, that is the main goal of the computational model. However, the resulting values of the other nodes are also shown: they are useful for further analysis, where thresholds for each feature are considered.

8. EXPERIMENTAL SETTING

Our experiment shows the choice between a doctor and a medical apparatus in the medical field. We assume that the choice is mainly driven by trustfulness. We have considered two situations: a "Routine Visit" and an "Emergency Visit". We have built four FCMs representing trustfulness for doctors and machines in those two situations. Even if the structure of the nets is always the same, the values of the nodes and the weights of the edges change in order to reflect the different situations. For example, in the "Routine Visit" scenario, Ability has a great causal power, while in the "Emergency Visit" one the most important factor is Accessibility. It is also possible to alter some values in order to reflect the impact of different trustier personalities in the choice. For example, somebody who is very concerned with Danger can set its causal power to *very high* even in the "Routine Visit" scenario, where its importance is generally low. In the present work we do not consider those additional factors; however, they can be easily added without modifying the computational framework.

8.1 Routine Visit Scenario

The first scenario represents many possible routine visits; there is the choice between a doctor and a medical apparatus. In this scenario we have set the initial values (i.e. the beliefs sources) for the Doctor hypothesizing some direct experience and common sense beliefs about doctors and the environment.

Most values are set to zero; the others are:

- Ability – Direct Experience: quite (+0.3);

- Ability – Categorization: very (+0.7);

- Accessibility – categorization: quite negative (-0.3);

- Unharmfulness – categorization: some negative (-0.2);

- Opportunity – Reasoning: some (+ 0.2);

- Danger – Reasoning: some negative (-0.2).

For the machine we have hypothesized no direct experience. These are the values:

- Efficacy – Categorization: good (+0.6);

- Accessibility – Categorization: good (+0.6);

- Unharmfulness – Categorization: quite negative (- 0.3);

- Opportunity – Reasoning: some (+0.2);

- Danger – Categorization: quite negative (- 0.3);

- Danger – Reasoning: quite negative (-0.3).

We have also considered the causal power of each feature. These values are the same both for the Doctor and the Machine. Most values are set to mildly relevant (+0.5); the others are:

- Ability: total causation (+1);

- Accessibility: only little causation (+0.1);

- Unharmfulness: middle negative causation (-0.4);

- Opportunity: only little causation (+0.1);

- Danger: little negative causation (-0.2).

The results of this FCM are shown in Figure 4: Trustfulness for the Doctor results good (+0.57) while trustfulness for the machine results only almost good (+0.22).

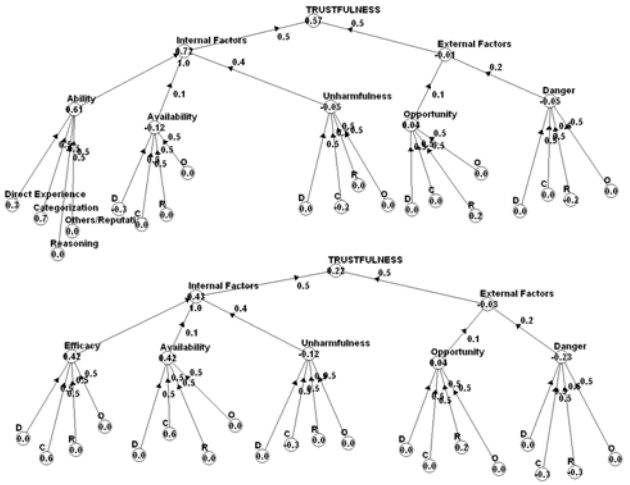

Figure 4. Routine Visit FCMs for the Doctor (top) and the Machine (bottom)

The FCMs are quite stable with respect to minor value changes; setting Machine's Accessibility – Direct Experience to good (+0.6), Accessibility – Categorization to really good (+0.8) and Danger – Categorization to little danger (-0.5) results in a non dramatic change in the final value, that changes from *almost good (+0.23)* to *quite good (+0.47)* but does not overcome the Doctor's Trustfulness. This is mainly due to the high causal power of Ability with respect to the other features. We can also see the influence of different personalities. For example, if we assume that Doctors are supposed to involve high external risks (Danger – Reputation: +1), with the usual values, the trustier's Trustfulness does not change very much (*good (+0.47)*). But if the patient is somebody who gives high importance to Danger (danger: *total causality (-1)*), the Doctor's Trustfulness decreases to *negative (-0.42)*.

8.2 Emergency Visit Scenario

We have hypothesized an emergency situation where somebody needs a quick visit for an easy task (e.g. a injection). In this scenario the values for the nodes are the same as before, but some edges drastically change: Reliability becomes very important and Ability much less. The values for the edges are:

- Ability: little causation (+0.2);

- Willingness: very strong causation (+1);

[2] We have used a slightly modified implementation of the Fuzzy Cognitive Map Modeler described in [12].

- Unharmfulness: strong negative causation (-0.8);

- Opportunity: middle causation (+0.5);

- Danger: quite strong causation (+0.6).

The results also change drastically: Trustfulness for the Doctor is *only slightly positive (+0.02)* and for the Machine it is *quite good (+0.29)*. The FCMs are very stable; altering some settings for the Doctor (Ability – Direct Experience: *very good* and Danger – Categorization: *only little danger*) results in a change in the Trustfulness value that become *almost good* but does not overcome the Machine's one. We obtain the same results if we suppose that Doctor's Ability - Direct Experience: *perfect* and Ability's Causal Power: *very strong*. On the contrary, if we introduce a big danger (+1) either internal (*harmfulness*) or external (*danger*) in each FCM the trustfulness values fall to *negative* in both cases (respectively -0.59 and -0.74 for the doctor; and -0.52 and -0.67 for the machine).

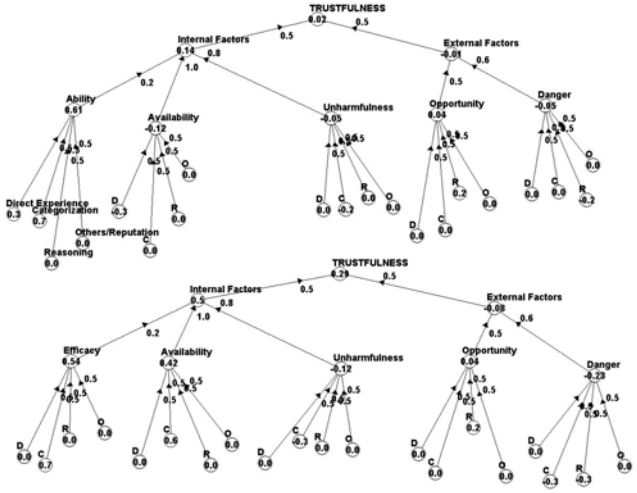

Figure 5. Emergency Visit FCMs for the Doctor (top) and the Machine (bottom)

9. TRUSTFULNESS AND DECISION

Obtaining the Trustfulness values is only the first step. In order to make the final choice (e.g. between a doctor and a machine in our scenarios) we have to take into account other factors, mainly costs and possible saturation thresholds for the various features. The main factor is represented by costs: we can consider the decision (for example to decide between x or y) as a standard costs-benefits product. The cost can even represent a threshold for the decision (e.g.: *I trust him but cost is too high*). However, there are other factors involved. FCMs not only show the overall Trustfulness value, but they show the values of all the factors that contribute to the final trust value: ability, reliance, danger, etc. We can fix a threshold for one or more features and inhibit a choice even if Trustfulness is acceptable (e.g.: *I trust him but risk is too high*). It is important to distinguish between the trust evaluation and the final decision; we can make our evaluation even if there are some factors that constrain the decision; for example we can evaluate the quality-price ratio of something even if we know we have not enough money to buy it. An FCM alone cannot take into account *necessary conditions* for a decision: this result can be reached only adding additional thresholds. This seems to us a good cognitive feature: even if there is an evidently unsatisfied *necessary condition*, normally humans do compute a trust value; they take into account this factor only as a second step (in the decision process). At the same way, we can judge something trustable but decide not to trust it. For example, in the game of poker, we can judge that a bet is a good action, and even do it if not so much money is involved; but we can decide not to do it if too much money is involved (the risk of loosing the money is too high for us). In this case we do not think that it is not a good action; we judge the bet a good action to do, but we simply could not take the risk.

10. EXPERIMENT DISCUSSION

The two scenarios try to take into account all the relevant factors for Trustfulness: beliefs sources, basic beliefs and their causal power. Moreover, FCMs allow to experiment changes in values due to different personalities. As already specified, belief sources are figured values, possibly derived from inner FCMs where many beliefs play their role. We have assumed four types of beliefs sources, but for many of them we give no values. We have set all their causal power to *middle causality* (+0.5) in order to let them be "neutral" in the experiments. Some different personalities can augment or reduce the values (e.g.: somebody who cares only about his own experience may assign a strong causal power to the corresponding edges). Basic beliefs, both internal and external, are the core of the analysis; we have expanded the original model [1, 2] by representing and quantifying the different importance of trust components/determinants (for different personalities or different situations). Our experiments show that the relative importance assigned to each feature may drastically change the results. Most of the differences in FCM's behavior is due to the strong causal power assigned to Ability (Routine Visit scenario) and accessibility (Emergency Visit scenario), even if the Basic Beliefs values are the same.

10.1 Evaluating the Behavior of the FCMs

We conducted several experiments modifying some minor and major beliefs sources in the FCM of Routine Visit Scenario for the Doctor. This allows us to evaluate their impact for the overall results. In the normal FCM the Trustfulness value is .57.

Table 1. Data for some minor factors (e.g. Unharmfulness)

Modified Factors	Old Value	New Value
Unharmfulness – Categorization from .2 to .3	.57	.57
Unharmfulness – Categorization from .2 to .4	.57	.55

Table 2. Data for some mayor factors (e.g. Ability)

Modified Factors	Old Value	New Value
Ability – Direct Experience from .3 to .2	.57	.52
Ability – Direct Experience from .3 to .1	.57	.45
Ability – Categorization from .7 to .65	.57	.48
Ability – Categorization from .7 to .75	.57	.66
Ability – Categorization from .7 to .8	.57	.71

We can see that the FCMs are quite stable: changing minor factors does not lead to catastrophic results. However, modifying the values of some mayor factors can lead to significant modifications; it is very important to have a set coherent parameters and to select very accurately the most important factors. However, our first aim is not to obtain an exact value for trustfulness for each FCM; on the contrary, even if we consider the whole system a qualitative approach, it has to be useful in order to make comparisons among competitors (i.e. the Doctor and the Machine in our scenarios). So, an important question about our system is: how much can I change the values (make errors in evaluations) and conserve the advantage of a competitor over the other? In the Routine Visit Scenario the two Trustfulness values are far one from another (.57 for the Doctor vs. .23 for the Machine). Even if we change several factors in the Machine's FCM (all .6 become .7 and all .2 and .3 become .4) its Trustfulness become .46 and do not overcome its competitor's one.

10.2 Personality Factors

Given the way in which the network is designed, it is clear that the weights of the edges and some parameters of the functions for evaluating the values of the nodes are directly expressing some of the personality factors. It is true that some of these weights should be learned on the basis of the experience (for now we do not consider the learning process). On the other hand, some other weights or structural behaviours of the network (given by the integrating functions) should be directly connected with personality factors. For example, somebody who particularly cares about his safety can overestimate the impact of danger and unharmfulness, or even impose a threshold for the final decision. Each personality factor can lead to different trust values even with the same set of initial values for the beliefs sources. Many personalities are possible, each with its consequences for the FCM; for example: *Prudent*: high danger and unharmfulness impact; *Too Prudent*: high danger and unharmfulness impact, additional threshold on danger and unharmfulness for decision; *Auto*: high direct experience impact, low impact for the other beliefs sources; *Focused on Reputation*: high reputation impact, low impact for the other beliefs sources.

Some personality factors imply emotional components, too. They can lead to important modifications of the dynamics of the FCM, for example modifying the choice of the heuristic for combining homogenous and heterogeneous fonts. In this paper we do not present the additional experiments we are doing using the personality factors.

11. CONCLUSIONS

Our experiments aim to describe the dynamics of trust and to capture its variations due to beliefs sources variation, and the different importance given to the causal links and personality factors. The scenarios presented here fail to capture many factors; in addition, we have assigned values and weights more as a matter of taste than by experimental results. More, the results of the experiments are shown as an attempt to describe the behavior of this kind of system; for example, its additive properties or the consequences of the choice of the threshold function. The adequacy of such a behavior to describe cognitive phenomena is an open problem. However, the experimental results show that it is possible to mimicry many commonsense assumptions about how trust varies while some features are altered; our aim was in fact to capture trust variations more than assign absolute values to it. In our view, this experiment confirms the importance of an analytic approach to trust and of its determinants, not simply reduced to a single and obscure probability measure or to some sort of reinforcement learning. Our future work will focus on building the belief fonts values starting from the single believes (splitting the contribute of values and credibility measures); at the same time we want to extend the architecture in order to take into account some personality factors (able to change the impact of some factors); we plan to maintain the same computational framework. Obtaining a value for Trustfulness represents only to the first step of the process of assigning trust; in order to make an effective decision (to trust or not to trust) several other factors are involved: mainly costs and thresholds over some specific features (sometimes determined according to personality factors, too).

12. REFERENCES

[1] Castelfranchi, C; Falcone, R., Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In Proceedings of the Third International Conference on Multi-Agent Systems, pages 72-79, Paris France, 1998.

[2] Falcone R., Castelfranchi C., (2001). Social Trust: A Cognitive Approach, in *Trust and Deception in Virtual Societies* by Castelfranchi C. and Yao-Hua Tan (eds), Kluwer Academic Publishers, pp. 55-90.

[3] Kosko, B. Fuzzy Cognitive Maps. International Journal Man-Machine Studies, vol. 24, pp.65-75, 1986.

[4] http://www.iis.ee.ic.ac.uk/~alfebiite/ab-home.htm

[5] J Carbonell: Towards a process model of human personality traits. *Artificial Intelligence, 15,1980.*

[6] Castelfranchi C., de Rosis F., Falcone R., Pizzutilo S., (1998) Personality traits and social attitudes in Multi-Agent Cooperation, *Applied Artificial Intelligence Journal*., special issue on "Socially Intelligent Agents", n. 7/8, vol.12, pp. 649-676.

[7] C Elliott: Research problems in the use of a shallow Artificial Intelligence model of personality and emotions. *Proceedings of the 12th AAAI*, 1994.

[8] C. Jonker and J. Treur (1999), Formal Analysis of Models for the Dynamics of Trust based on Experiences, Autonomous Agents '99 Workshop on "Deception, Fraud and Trust in Agent Societies", Seattle, USA, May 1, pp.81-94.

[9] M. Schillo, P. Funk, and M. Rovatsos (1999), Who can you trust: Dealing with deception, Autonomous Agents '99 Workshop on "Deception, Fraud and Trust in Agent Societies", Seattle, USA, May 1.

[10] Dragoni,A. F., 1992. A Model for Belief Revision in a Multi-Agent Environment. In *Decentralized AI - 3,* Y. Demazeau, E. Werner (eds), 215-31. Amsterdam: Elsevier.

[11] Castelfranchi, C. 1996. Reasons: Belief Support and Goal Dynamics. *Mathware & Soft Computing, 3*. 1996, pp. 233-47.

[12] http://www.users.voicenet.com/~smohr/FCMApplication.htm