11

A Fuzzy Implementation for the Socio-Cognitive Approach to Trust

In this chapter[1] we will show a possible implementation of the socio-cognitive model of trust developed in the other chapters of the book. This implementation (Falcone *et al.,* 2005) uses a fuzzy approach (in particular, it uses the so-called Fuzzy Cognitive Maps –FCM (Kosko, 1986). In particular our attempt is to show, using a specific implementation, how relevant a trust model is based on beliefs and their credibility.

As previously described, our model introduced a degree of trust instead of a simple probability factor since it permits trustfulness to be evaluated in a rational way: Trust can be said to consist of, or even better (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. In any case this probability index is based on (derives from) those beliefs and evaluations. In other words, the global, final probability of the realization 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 its willingness, persistence, engagement, competence: *internal attributions*) and the probability of having the appropriate conditions (opportunities and resources: *external attributions*) for the performance and for its success, and of not having interferences and adversities (*external attributions*).

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 significantly modify the final trust.

[1] We would like to thank Giovanni Pezzulo for his precious contribution on this chapter.

Trust Theory: A Socio-Cognitive and Computational Model Cristiano Castelfranchi and Rino Falcone
 ISBN: 978-0-470-02875-9

11.1 Using a Fuzzy Approach

Given our purpose of modelling a graded phenomenon like trust (that is difficult to estimate experimentally) we have chosen a Fuzzy Logic Approach (FLA). A clear advantage with FLA is the possibility of using natural language labels (like: '*this doctor is very skilled*') to represent a specific real situation. In this way, it is more direct and simple to use intervals rather than exact values.

In addition, the behavior of these systems (e.g. their combinatorial properties) seems to be good at modeling several cognitive dynamics (Dubois and Prade, 1980), even if finding 'the real function' for a mental operation and estimating the contribution of convergent and divergent belief sources remain ongoing problems.

We have used an implementation based on a special kind of fuzzy system called Fuzzy Cognitive Maps (FCM); they allow the value of the trustfulness to be computed, 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 over one another.

11.2 Scenarios

The scenario we are going to study is *medical house assistance* and we will look at it in two particular instances:

a) A *doctor* (a human operator) visiting a patient at home, and
b) A *medical automatic system* used to support the patient (without direct human intervention).

The case studies under analysis are:

- An *emergency situation*, in which there is a need to identify what is happening (for example, a heart 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 (perhaps just an injection).
- A *routine situation*, in which there is a systematic and specialist therapy which needs to be applied (using quite a complex procedure) but in which there is no immediate danger to cope with.

We will show how the following factors can produce the final trust for each possible trustee who is dependent on it:

- 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 factor that we are able to characterize some *personality traits* of the agents (Castelfranchi *et al.*, 1998).

11.3 Belief Sources

As shown in Chapter 2, our trust model is essentially based on specific beliefs which are both a *basis* of trust and also its *sub-components* or *parts*. These 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 (see Chapter 3): *the quantitative dimensions of trust are based on the quantitative dimensions of its cognitive constituents*.

However, what is the origin and the justification of the strength of beliefs? Our answer is: Just their sources. In our model, 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 (Jonker & Treur, 1999), (Schilloet *et al.*, 1999). But they only consider direct interaction (experience) or reputation as sources. In this implementation we have considered four possible types of belief sources:

- *direct experience* (how the personal – positive or negative – experience of the trustor 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 trustor beliefs).

We do not consider learning in the model's dynamic. We are just modeling the resulting effects that a set of trustor's basic beliefs (based on various sources) have on the final evaluation of the trustee's trusfulness about a given task and in a specific situation. At present we do not consider how these effects feed back on the basic beliefs.

11.4 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.

For each kind of source we have to consider the impact it produces on trustor's beliefs about trustee's features. These impacts result from the composition of the value of the content (property) of that specific belief (the belief's object) 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;
- *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 *trustor evaluates this source* (with respect to this belief).

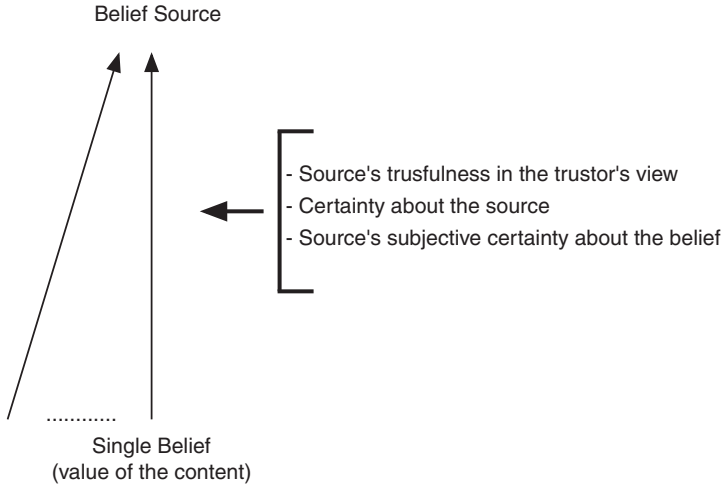

Figure 11.1 From single beliefs to the belief source

Those beliefs are not all at the same level. 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 *Figure* 11.1); 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 trustor'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 trustor. At this level all the contributions of the various sources of the same type are integrated.

Let us consider as an example the belief source of the kind '*Reputation*' about a *doctor's ability* (see *Figure* 11.2). In order to have a value, we have to consider many opinions about

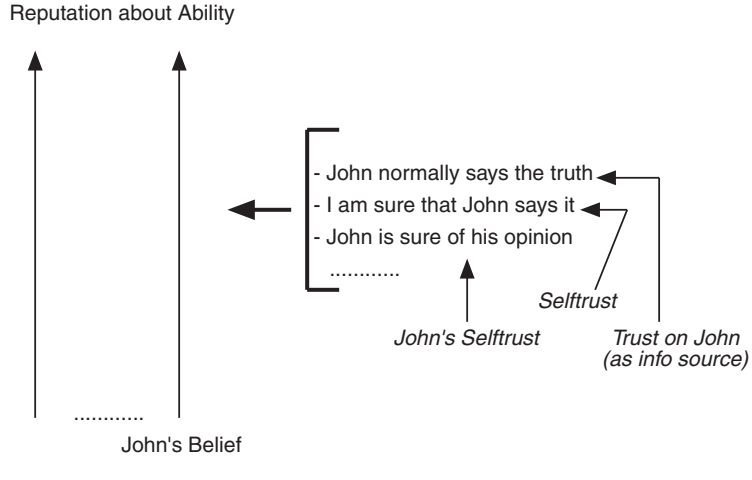

Figure 11.2 Case of belief source of Reputation

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's content: '*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 at his work*);
- the degree of certainty that the trustor has about the fact that John has expressed this opinion (*I am sure* that John told me (thinks) that, etc.);
- how good John considers his own belief (when John says: 'I think', he could mean: *I am sure/ I am quite sure/ I am not so sure* and so on);
- the credibility of John's opinion (from the trustor's point of view).

The first factor represents 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 trustor's mental world. The second factor represents a trustor's degree of certainty that the source expressed (communicated) that belief (it is also linked with the trustor's self-trust). The third factor represents an epistemic evaluation that the source makes on the communicated belief. Finally, the fourth factor represents a degree of trust in the source's opinion, and it depends on a set of trustor's beliefs about source's credibility, ability to judge and so on.

The second, third and the fourth factor 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.

The impact factors are not evaluation beliefs, but rather epistemic ones: they describe the way to see the other beliefs and their degree of certainty. So, at the level of building belief sources, evaluation and epistemic factors are separated; from the belief sources level up, in our FCM representation, they are combined in a unique numerical value.[2]

11.4.1 A Note on Self-Trust

Self-trust is a belief that relies on many beliefs, as, in general, trust is: their belief-FCM can be built in the same manner. As for trustfulness, self-trust is specific of a task or of a context. Among belief sources there can be, as usual, personal opinions and others' ones – i.e. reputation.

In self-trust computation there is also a set of motivational factors: self-image, auto-deceiving, and so on. Since our implementation does not represent motivational factors, at this moment we are not able to take into account these factors; so we calculate self-trust in the same way trust is calculated.

[2] Even quantitative information (how much I know about) is combined; for example, a low value about the ability of a doctor can derive from: low evaluation; low confidence in my information sources, little information.

11.5 Implementation with Nested FCMs

In order to understand the following parts of the work, we need to describe how a belief source value is computed starting from many different opinions[3] of different sources (e.g. in a MAS system, each agent is a source and can communicate an opinion about something). In an FCM this situation is modelled with a set of nodes representing single beliefs that have an edge ending on the (final) belief source node of the FCM. For each of these nodes, two values are relevant: the value of the node itself and the value of the edge.

The value of the node, as usual in this kind of model, corresponds directly to the fuzzy label of the belief; for example, *John says that the doctor is quite good at his work* can be considered as a belief about the doctor's ability with value 0.5 that impacts over the others/reputation belief source of a doctor's ability.[4]

Computing the *impact factor* of this belief (i.e. the value of the edge in the FCM) is more difficult. We claim that the impact represents not a cognitive primitive; rather, it has to be computed by a *nested FCM*, that takes into account mainly epistemic elements about the opinion itself and its source.

In our experiments with FCMs evaluation and epistemic issues are mixed up in a single value; this was a methodological choice, because we wanted to obtain one single final value for trustfulness. But this is the place where the two different kinds of information can be kept separate because they have a different role. Figure 11.3 shows many elements involved in this FCM: mainly beliefs about the source of the belief, grouped into three main epistemic features. Here we give an example of such an FCM in the medical domain. This FCM has single beliefs that impact on these features; the resulting value represents the final impact of a single belief over the belief source node.

This nested FCM was filled in with many nodes in order to show the richness of the elements that can intervene in the analysis. A similar FCM can be built for each single belief that impacts into the belief sources nodes; some of those nested FCMs have overlapping nodes, but in general each belief can have a different impact, depending on epistemic considerations.

It is possible to assign different impacts to the three different epistemic features; in this case we wanted to give them the same importance, but it depends from both the contingent situation, from personality factors and even from trust: for example, my own opinions can be tuned by self-trust (e.g. sureness about my senses and my understanding ability), and Mary's opinions can be tuned by trust about Mary. This leads to a very complex structure that involves trust analysis about all the sources (and about the sources' opinions about the other sources). For the sake of simplicity in the example we use all maximal values for impacts.

In general, it is important to notice that the 'flat' heuristic (same weights) we use in order to mix the different factors is not a cognitive claim, but a need derived from simplicity and lack of empirical data. In the following paragraph we investigate a very similar problem that pertains to how to sum up the different belief sources.

[3] We call this information *opinions* and not *beliefs* because they are not into the knowledge structure of an agent; an agent can only have a belief about another agent's beliefs (*John says that the doctor is good* is a belief of mine, not of John). This belief sharing process is mediated by opinions referred by John, but it can even be false, misleading or misinterpreted. What is important, however, is that beliefs are in the agent's cognitive structure, whether they correspond or not to other agent's opinions, beliefs or even to reality.

[4] It is important to notice that this node does not represent an opinion of John; it represents a belief of the evaluator, that can be very different from the original John's opinion (for example, it can derive from a misunderstanding).

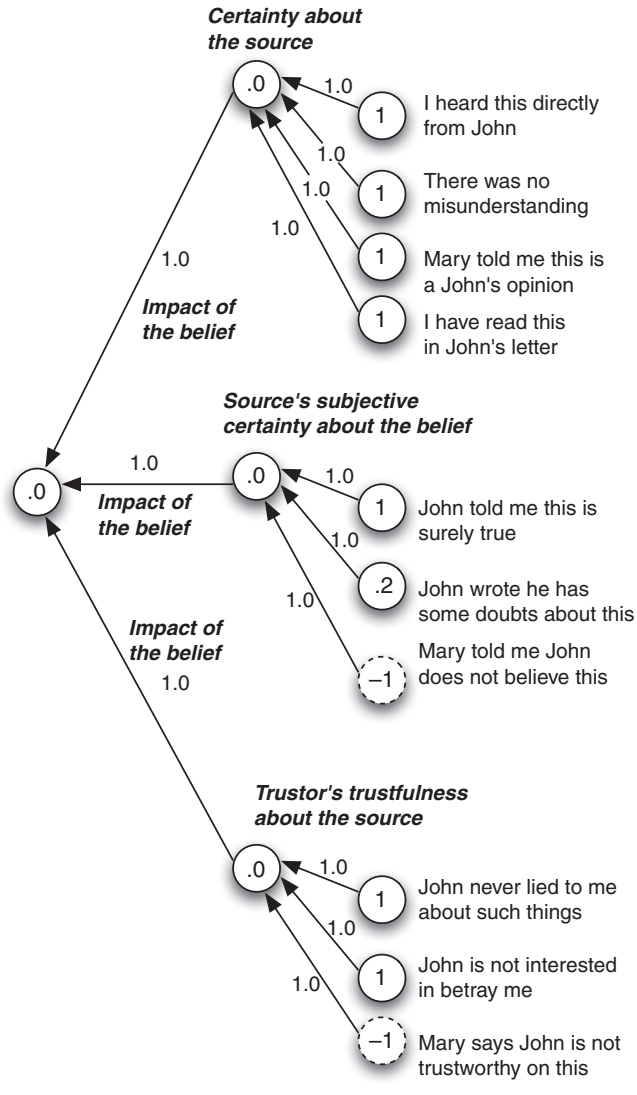

Figure 11.3 A Nested FCM

11.6 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 (Dragoni, 1992), (Castelfranchi, 1996). It is in particular an evident problem in the case in which we are going to model human behaviors. 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 simply

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 Dr Smith is not too good a doctor and also not too bad a doctor while two other persons give us two diverging evaluations on Dr 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 Dr White and Dr Smith, and our decision would be guided by other criteria. *These criteria are linked with context, emotions, personality factors*. We could have people who, in the presence of diverging opinions, decide to suspend judgment (they become unable to decide), or people who take into consideration the best opinion (*optimistic personality*), or, on the contrary, people who take into 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 the object of empirical analysis or simulations. The model itself is independent to those heuristics, that is they can be easily substituted.

11.7 Homogeneous 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 more relevant, our own personal experience or the reputation about a specific ability of a person? There is not a definitive 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.

11.8 Modeling Beliefs and Sources

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

Table 11.1 Internal and external factors for the *automated medical system*

Internal factors	***Ability or Competence beliefs***	They 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
	Disposition or Availability beliefs	They are linked to the reliability of the machine, its regular functioning, its ease of use; possibly, its adaptability to new and unpredictable uses
	Unharmfulness beliefs	They concern the lack of internal/ intrinsic risks of the machine: the dangers implied in the use of that machine (for example side effects for the trustor's health), the possibility of breaking and so on
External factors	***Opportunity beliefs***	Concerning the opportunity of using the machine, independently 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 who might support its use, etc.)
	Danger beliefs	They are connected with the absence of the systemic risks and dangers external to the machine that could harm the user: consider for example the risk for the trustor'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

What are the meanings of our basic beliefs in the case of the doctor and in the case of the automated medical system? For both the latter and former, the internal and external factors are shown in Table 11.1 and 11.2.

Each of the above mentioned beliefs may be generated through 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.

11.9 Overview of the Implementation

An FCM is an additive fuzzy system with feedback; it is well suited to the representation of 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.

Table 11.2 Internal and external factors for the *doctor*

Internal factors	***Ability or Competence beliefs***	They concern the (physical and mental) skills of the doctor; his/her ability to make a diagnosis and to solve problems
	Disposition or Availability beliefs	They 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).
	Unharmfulness beliefs	They concern the absence (lack) of the risk of being treated by a doctor; namely the dangers of a wrong diagnosis or intervention (for example, for the health of the trustor).
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 trustor is very close to a hospital in which there is an efficient service of fast intervention; or again, even if the trustor 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 (with heavy traffic, frequent strikes) that could hamper the mobility of the doctors and of their immediate transfer to the site where the patient is.
	Danger beliefs	These beliefs concern 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 trustor's house is poor and not too clean, the trustor could see the visit of a person (the doctor in this case) as a risk for his/her reputation.

In order to design the FCM and to assign a value to its nodes we need to answer four questions:

1) Which value do I assign to this concept?
2) How sure am I of my assignment?
3) What are the reasons for my assignment?
4) How much does concept impact on another linked concept?

We address the *first and the second question* above by 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 trustor is *pretty sure* that the trustee is *rather good* or that he/she is *rather sure* that the trustee is *really excellent*, etc. In this implementation we do not address how the degree of credibility/certainty of the belief combines with the degree of the content dimension (even if this analysis is quite relevant); we just use a single resulting measure.

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 by 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 basic trust model.

FCMs allow causal inference to be quantified 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 modeled as a (strong) positive causal inference between the two concepts of accessibility and trustfulness. FCMs also allow the influence of different causal relations to be summed up. For example, adding another statement: 'Doctors are *very good* in their ability, but this is a *minor factor* in an emergency situation' means adding a new input about 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.

11.9.1 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.

FCM is different from standard fuzzy techniques, in that it requires the use of crisp input values; we have used the average of the usual ranges, obtaining the following labels, both for positive and negative values: *quite*; *middle*; *good;* etc. However, as our experiments show, even with little variation of these values in the same range, the FCMs are stable and give similar results.

As Figure 11.4 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 kinds of 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,

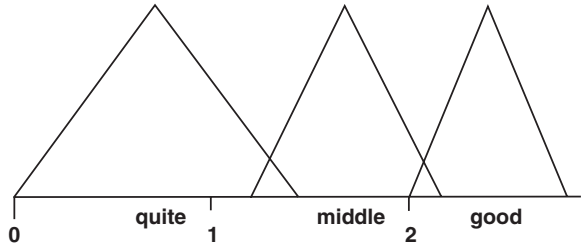

Figure 11.4 Fuzzy Intervals. (Reproduced with kind permission of Springer Science+Business Media © 2003)

changes in big (positive or negative) values have more impact on the FCM, this is a tolerable result even if it does not correspond with a general cognitive model.

11.10 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': *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 as 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, too.

11.11 Running the Model

Once the initial values for the first layer (i.e. belief sources) are set, the FCM starts running. 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.

11.12 Experimental Setting

Our experiments show 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 factors is *Accessibility*.

It is also possible to alter some values in order to reflect the impact of *different trustor 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.

11.12.1 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)*;
- Avaialability – 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)*

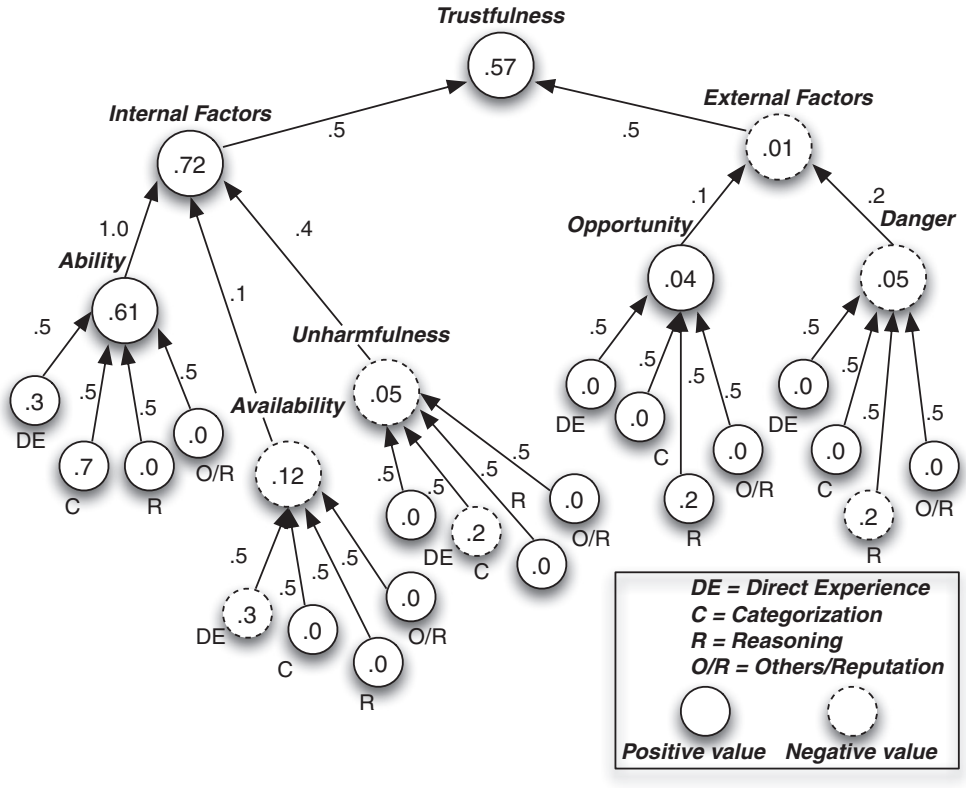

Figure 11.5 Routine Visit FCMs for the Doctor. (Reproduced with kind permission of Springer Science+Business Media © 2003)

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 11.5 and 11.6: trustfulness for the doctor results *good (+0.57)* while trustfulness for the machine results only *almost good (+0.22)*.

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.

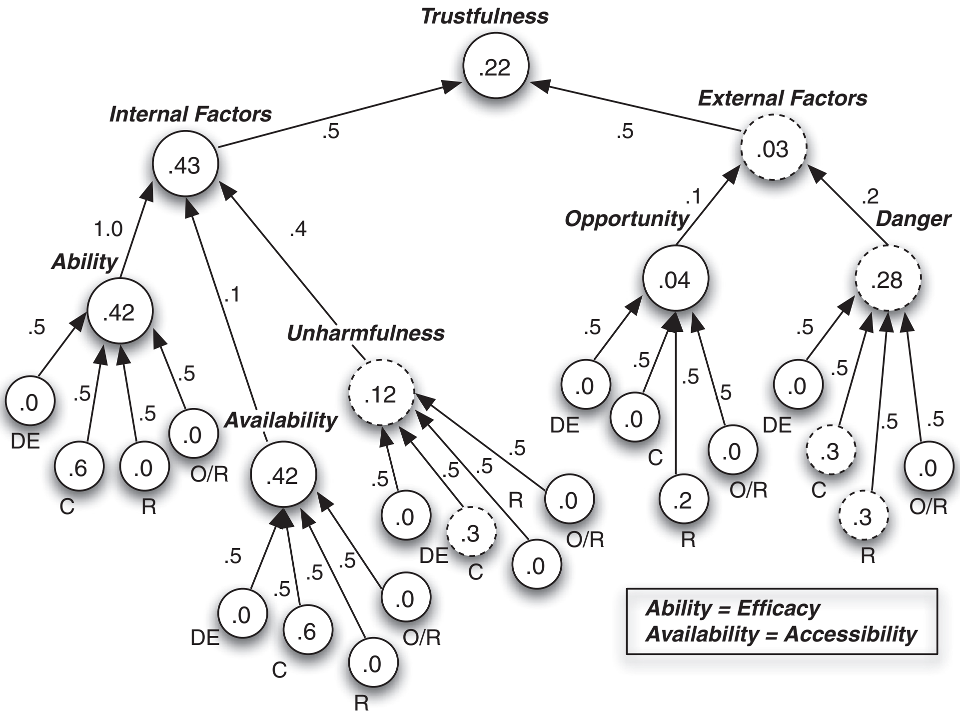

Figure 11.6 Routine Visit FCMs for the Machine. (Reproduced with kind permission of Springer Science+Business Media © 2003)

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 trustor'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)*.

11.12.2 Emergency Visit Scenario

We have here 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);
- Unharmfulness: strong negative causation (−0.8);
- Opportunity: middle causation (+0.5);
- Danger: quite strong causation (+0.6).

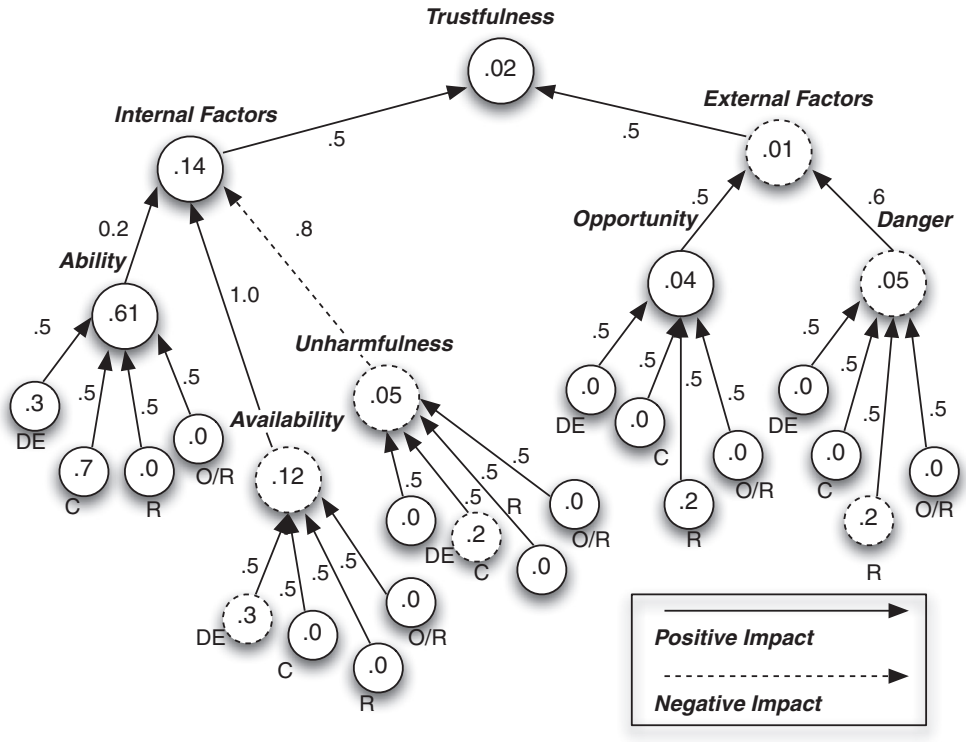

Figure 11.7 Emergency Visit FCMs for the Doctor. (Reproduced with kind permission of Springer Science+Business Media © 2003)

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)* (see Figure 11.7 and 11.8).

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 becomes *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).

11.12.3 Trustfulness and Decision

We consider three steps: evaluation (i.e. how much trust do I have); decision (to assign or not assign a task); delegation (make the decision operative). 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.

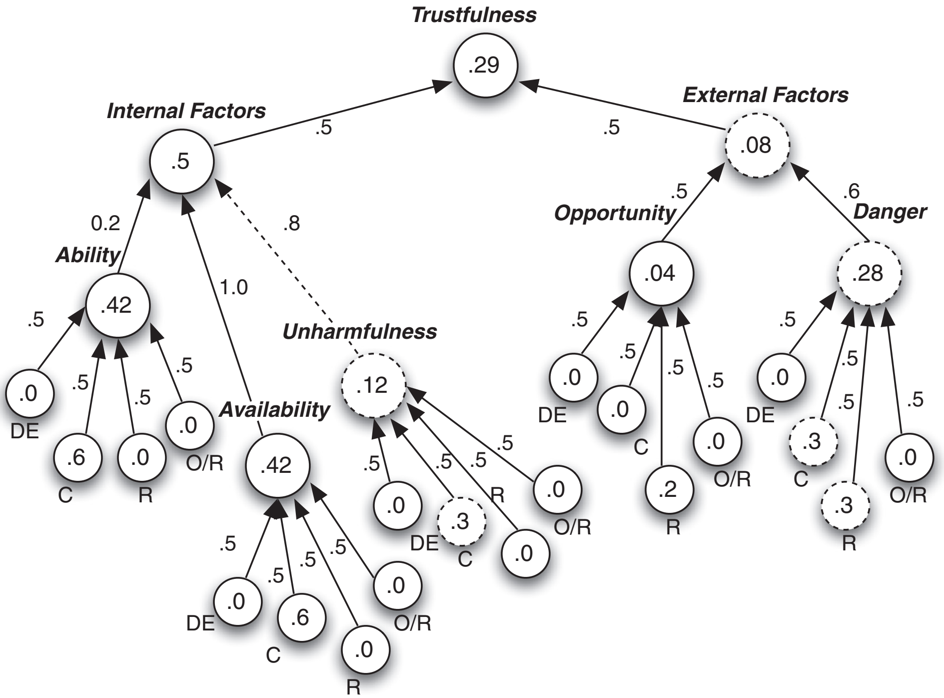

Figure 11.8 Emergence Visit FCMs for the Machine. (Reproduced with kind permission of Springer Science+Business Media © 2003)

FCMs not only show the overall trustfulness value, but also the values of each belief. We can fix a threshold for one or more features and inhibit a choice even if trustfulness is acceptable (i.e. 'I trust him, but the danger is too high'). In addition, the final function for decision has to also take into account the costs for each decision choice. In the present analysis we do not consider here these additional factors.

11.12.4 Experimental 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 experimentation of 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 our model (see Chapters 2 and 3 in this book) 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 are 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.

11.12.5 Evaluating the Behavior of the FCMs

We have 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 on the overall results.

We can see that the FCMs are quite stable: changing minor factors does not lead to catastrophic results. However, modifying the values of some major factors can lead to significant modifications; it is very important to have a set of coherent parameters and to select the most important factors very accurately.

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 as 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 removed from one another (0.57 for the doctor vs. 0.23 for the machine). Even if we change several factors in the machine's FCM its trustfulness does not overcome its competitor's one.

11.12.6 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. 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 their safety can overestimate the impact of danger and unharmfulness, or even impose a threshold on 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.

To summarise, we can say that our experiments aim to describe the dynamics of trust and to capture its variations due to belief 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 through

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 ongoing problem.

However, the experimental results show that it is possible to mimic 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.

In the next two paragraphs we introduce:

- some learning mechanisms with respect to the reliability of the belief sources; and
- some comparison experiments among different strategies for trusting other agents using a *Contract Net* protocol: we will show how our cognitive approach presents some advantages.

11.13 Learning Mechanisms

In the previous paragraphs we have considered the belief sources as a static knowledge of the agents. Briefly in this part we show how it could be possible to extend this approach by trying to model some of the dynamics generated by a learning process and by trust itself.

We give an agent in a MAS system the capacity to evaluate its 'sources of opinions', i.e. the other agents, according to a specific advising attitude parameter: *trust in Y as an information source*, (that is different from trust in *Y* to perform differently).

In order to build a belief source, we considered a node representing a single belief and an edge representing the impact of this single belief: the value of the edge (the impact factor) represents *the validity of the source* with respect to this single communicative episode. Some elements of this episode are unique (e.g. certainty about the source) but others are shared between all the other communicative episodes with the same source: the trustfulness of the source is applicable to all the class of the possible beliefs about the opinions of a source about an argument. These values can be learned and applied to future cases; for example, if it results, from some interactions with John, that he systematically lies, the impact of (my belief about) his opinion (e.g. *John says p*...) will drastically diminish or even become negative, and this value can be used in further interactions.

As shown the FCM computes its results until it stabilizes. This leads to a stable result for each node involved in the FCM. Here we propose a second phase: *the FCM 'evaluates its sources', i.e. modifies the impact of each single belief source according to the final value of its belief source node*.

For example, many nodes $n_1, \ldots, n_n$ representing single beliefs (opinions given by the source) can contribute to the value of a belief source node N. Each node n_j (with $1 \leq j \leq n$) has an impact $i_1, \ldots, i_n$ over N; the impact value is calculated by the inner FCM previously described. After the FCM stabilization, the difference between the (final) value of the belief source and the value of each single belief (of the source) can be seen in information terms as an *error*.

The *learning phase* consists in trying to minimize errors; in our terms, the impact of a bad opinion (and the importance of the corresponding source), has to be lowered; the reverse is

also true. In order to achieve this result, we have two different strategies: *implicit* and *explicit revision*.

11.13.1 Implicit Revision

In an *implicit revision* procedure, the *error* (i.e. the difference between N and n_j) can be back-propagated and modify the value of i_j. The rationale is that the FCM adjusts the evaluation of its sources, the impact of their opinions (as mediated by our beliefs); for example, in the case of a source that systematically lies (gives as output the opposite of the final value of the target node), step to step its impact will be nearer to *–1*.

The revision process has two steps. The first step is the change of the impact factor: a standard back-propagation algorithm can be used in order to achieve this result. The second step leads to a feedback of this revised value over the nested FCM: some low-weight edges are assumed that back-propagate the value until the nested FCM stabilizes. For example, if the impact factor value was lowered, all the nodes in the nested FCM will be lowered a bit, until it stabilizes.

This procedure has to be better explained. Since the value of the edge (the impact factor) represents the validity of the source, this value changes as I have a feedback between an opinion and my final belief. For example, in evaluating the ability of a doctor, a significant difference between a value furnished by a source and the final value I assume, means that the source was not totally valid. This can result from different reasons: the source is not trustworthy, a misunderstanding, poor information and so on. With regard to these problems the implicit revision strategy is blind: it revises the value of the impact successively to all the nodes in the nested FCM, without caring *what nodes* are responsible for the error and so need to be changed.

This process adjusts the evaluation of a single belief, but it can be in part shared with other belief impact evaluations with respect to the same source: some nodes of the nested FCM apply only to the current situation (e.g. certainty about the source) but others are related to all the interactions with the same source (e.g. trustfulness about the source). So, this form of learning from a single episode generalizes for the following episodes. This kind of learning is non specific. We cansider a better one: explicit revision.

11.13.2 Explicit Revision

The *explicit revision* consists of the revision of some beliefs about the source; since these beliefs are part of the *inner FCM*, the (indirect) result is a modification of the impact. So, *explicit revision* means revising the values of some nodes of the inner FCM (or building new ones); the revised *inner FCM* computes a different impact value f.

In order to obtain explicit revision, the first important issue is to decide *where* to operate in the inner FCM. In some cases it would be useful to insert a new node representing a bad or good 'past experience' under the 'trustfulness' feature; in this case the value is easily set according to the usual set of fuzzy labels; an example of such a node can be *This time John was untrustworthy*.

Even if it is unrealistic to think that a single revision strategy is universal, there can be many heuristics: for example, a wrong opinion can be evaluated in different ways if I am sure

Table 11.3 Source evaluation

Willingness Episodes	The source reveals itself not to be trustworthy (*I know John intentionally lied to me*)
Competence Episodes	The source reveals itself not to be competent (*John did not lie but he was not able to give an useful information*)
Self trust Episodes	The evaluator was responsible for a misunderstanding (*I misunderstood*)
Accidental Problems	There were contingent problems and no revision is necessary (*there was an unpredictable circumstance*)

that somebody intended exactly this (there was no misunderstanding) or if I do not remember exactly what he said (or even if I am not sure that it was his opinion). In the latter cases, if I am not certain of the source, it can be better to assign the error to my evaluation rather than to ignorance of the source or even worse to his intention to deceive me. Such a change has no impact on other interactions with the same sources (but it can lead to change my self trust value).

11.13.3 A Taxonomy of Possible Revisions

There are many possible ways to evaluate an episode of interaction in order to learn from it and to decide to change one's beliefs. As we have shown, not only the sources' opinions, but also the full set of interaction episodes have to be categorized; from this kind of categorization the following belief revision process depends. For example, in order to comprehend the motivation of the discrepancy between the source evaluation and my evaluation ('*John says that this doctor is pretty good, but it results to me to be not so good...*'). The first thing to consider is 'what was the main factor' from which this discrepancy depends.

Obviously this decision process pertains to a cognitive apparatus and it is impossible at a pure-belief level; so a cognitive agent needs some revision strategies that individuate the error source and try to minimize it for the future. In Table 11.3 we propose a crude taxonomy of the problems that can intervene in an episode of interaction.

Implicit revision performs better with regard to computational speed. However, explicit revision has many advantages. First of all, taking into account single cognitive components allows a better granularity; this can make the difference where fine-grained distinctions are needed, for example in order to distinguish between trust somebody as an information source and as a specialist, or to distinguish a deceiver from a not informed source. Also, a single belief can be shared among many different FCMs, so this operation leads to the generalization and reuse of the obtained results.

In general, explicit revision takes into account the single cognitive components of trust, and this feature is one of our main desiderata. We derive trust from its cognitive components, i.e. from single agent's beliefs. So it is better to store information learned by experience into the same representation form (i.e. beliefs) rather than using compounded values (as an impact is), in order to integrate them into the representational and reasoning system of the agent.[5]

[5] However, it has to be noticed that since we have a fine-grained distinction between different belief sources, even the implicit mechanism results in being sufficiently accurate and specific for many purposes, even if it loses part of

Keeping them separate can lead to building different graphs and applying different revision heuristics, too.

The process of evaluation of the sources is described as a step that follows the stabilization of the FCM. In computational terms, thanks to the characteristics of the FCMs, these feedback processes can even be made in a parallel way. We see this process as a cognitive updating of beliefs: the parallel options are better to model gradual opinion shifts, while the two phases division allows the mimicry of counterfactual phenomena and drastic after decision changes.

When is it good for the system to evaluate its sources? In our experiments we assume, in a conservative way, that the mechanism only starts after a decision; the rationale is that a decision taken is a sufficient condition to assume that the stabilization reached is sufficient; less conservative criteria are possible, of course: this choice can be considered an heuristic rather than a part in the way the system works.

An interesting 'side-effect' of this source evaluation mechanism is that revision has less effect on well established sources (i.e. agents that have many interactions with us); we are less inclined to revise the stronger ones, mainly for economic reasons. The process described takes into account all the past experiences, so introducing a new example (or a counterexample) has less impact on the case of many interactions.

11.14 Contract Nets for Evaluating Agent Trustworthiness[6]

In this paragraph we show a first significant set of results coming from a comparison among different strategies for trusting other agents using a contract net protocol (Smith, 1980). We introduced three classes of trustors: a *random trustor*, a *statistical trustor*, a *cognitive trustor*. All the simulations were performed and analyzed using the cognitive architecture AKIRA (Pezzullo & Calvi, 2004).

The results show the relevance of using a *cognitive representation* for a correct trust attribution. In fact, a cognitive trustor performs better than a statistical trustor even when it has only an approximate knowledge of the other agents' properties.

11.14.1 Experimental Setting

We implemented a contract net with a number of trustors who delegate and perform tasks in a variable environment. Each agent has to achieve a *set of tasks* and is defined by a set of features: *ability set*, *willingness*, *delegation strategy*.

- The *Task set* contains the tasks an agent has to achieve; it is able either to directly perform these tasks, or to delegate them to some other agent.
- The *Ability set* contains the information about the agent's skills for the different tasks: each agent has a single ability value for each possible task; it is a real number that ranges in *(0,*

the expressiveness of the explicit one. Sometimes the trade off between computational power and expressiveness can lead to the adoption of the implicit mechanism.

[6] We would like to thank Giovanni Pezzulo and Gianguglielmo Calvi for their relevant contribution to the implementation and analysis of the model discussed in this paragraph.

1). At the beginning of the experiment these values are randomly assigned to each agent on each possible task.
- The *Willingness* represents how much the agent will be involved in performing tasks (e.g. how many resources, will or amount of time it will use); this modulates the global performance of the agent in the sense that even a very skilled agent can fail if it does not use enough resources. Each agent has a single willingness value that is the same for all the tasks it tries to perform; it is a real number that ranges in *(0, 1)*.
- The *Delegation strategy* is the rule an agent uses for choosing which agent to delegate the task to (e.g. random, cognitive, statistical). It is the variable we want to control in the experiments for evaluating which trustor performs better.

Agents reside in an *environment* that changes and makes the tasks harder or simpler to perform. Changes are specific for each agent and for each task: in a given moment, some agents can be in a favorable environment for a given task, some others in an unfavorable one. For example, two different agents, performing the same task, could be differently influenced by the same environment; or, the same agent performing different tasks in the same environment could be differently influenced by it in performing the different tasks. Influences range in *(−1, 1)* for each agent for each task; they are fixed at random for each simulation. The environment changes randomly during the simulations: this simulates the fact that agents can move and the environment can change. However, for all experiments, if a task is delegated in an environment, it will be performed in the same one.

11.14.2 Delegation Strategies

In the contract net, on the basis of the offers of the other agents, each agent decides to whom to delegate (Castelfranchi and Falcone, 1998) depending on their delegation strategy. We have implemented a number of different agents, having different *delegation strategies*:

- a ***random trustor***: who randomly chooses the trustee to whom to delegate the task. This kind of trustor has no a priori knowledge about: the other agents, the environment in which they operate, their previous performances. There is no learning. This is used as a base line.
- a ***statistical trustor***: inspired by a number of works, including (Jonker and Treur, 1999), assigns a major role to learning from direct interaction. They build the trustworthiness of other agents only on the basis of their previous performances, without considering specific features of these agents and without considering the environment in which they performed. It is one of the most important cases of trust attribution; it uses the previous experience of each agent with the different trustees (failures and successes) by attributing to them a degree of trustworthiness that will be used to select the trustee in a future interaction. There is a training phase during which this kind of trustor learns the trustworthiness of each agent through a mean value of their performances (number of failures and successes) on the different tasks in the different environments; during the experimental phase the statistical trustor delegates the most trustful agent (and continues learning, too). There is no trustor's ability to distinguish how the properties of the trustee or the environment may influence the final performance.

- a ***cognitive trustor***: this kind of trustor takes into account both the specific features of the actual trustee and the impact of the environment on their performance. In this implementation there is no learning for this kind of agent but an a priori knowledge of the specific properties of the other agents and of the environment. It is clear that in a realistic model of this kind of agent, the a priori knowledge about both the internal properties of the trustees and the environmental impact on the global performance will not be perfect. We did not introduce a learning mechanism for this kind of agent (even if in Section 11.13 we discussed this problem and showed potential solutions) but we introduced different degrees of errors in the knowledge of the trustor that corrupted their perfect interpretation of the world. The cognitive model is built using Fuzzy Cognitive Maps. In particular, two special kind of agents will be analyzed:
- ***best ability trustor***: who chooses the agent with the best ability score.
- ***best willingness trustor***: who chooses the agent with the best willingness score.

These two kind of cognitive agents can be viewed as having different 'personalities'.

11.14.3 The Contract Net Structure

We have performed some experiments in a *turn world*, others in a *real time world*. In the turn world the sequence is always the same. The first agent (randomly chosen) posts their first task (*Who can perform the task* τ*?*) and they collect all the replies from the other agents (*I can perform the task* τ *in the environment w*). All data given from the offering agents are true (there is no deception) and in particular the cognitive trustors know the values of ability and willingness for each agent (as we will see later, with different approximations).

Depending on their delegation strategy, the trustor delegates the task to one of the offering agents (in this case, even to themselves: self-delegation). The delegated agent tries to perform the task; if it is successful, the delegating agent gains one *Credit*; otherwise it gains none. The initiative passes to the second agent and so on, repeating the same schema for all the tasks for all the agents. At the end of each simulation, each agent has collected a number of *Credits* that correspond to the number of tasks that the delegated agents have successfully performed.

We have introduced no external costs or gains; we assumed that each delegation costs the same and the gain of each performed task is the same. Since the agents have the same structure and the same tasks to perform, gained credits are the measure of success of their delegation strategy.

In the *real time world* we have disabled the turn structure; the delegation script is the same, except for no explicit synchronization of operations. This means that another parameter was implicitly introduced: *time* to execute an operation. Collecting and analyzing messages has a time cost; agents who have more requests need more time in order to fulfill them. In the same way, agents who do more attempts in performing a task, as well as agents who reason more, spend more time. In *real time world* time optimization is another performance parameter (alternative or together with *credits*), and some alternative trust strategies become interesting: in real time experiments we introduced another strategy:

- the ***first trustful trustor***: it is a variant of the cognitive trustor and it has the same FCM structure; but it delegates to the first agent whose trust exceeds a certain threshold: this is

less accurate but saves the time of analyzing all the incoming messages. Furthermore, if some busy agent accepts only a limited number of tasks, or if there is a limited time span for performing the maximum number of tasks, it is important to be quick to delegate them.

11.14.4 Performing a Task

When an agent receives a delegation, it tries to perform the assigned task. Performing a task involves three elements: two are features of the agent (task specific ability and willingness). The third is the (possible) external influence of the environment. In order to be successful, an agent has to score a certain number of hits (e.g. *3*); a hit is scored if a random real number in *(0, 1)* is rolled that is less than its ability score. The agent has a number of tries that is equal to ten times its willingness value, rounded up (i.e. from 1 to 10 essays). The *environment can interfere* with an agent's activity giving a positive or negative modifier to each roll (so it interferes with ability but not with willingness). If the number of scored hits is sufficient, the task is performed; otherwise it is not.

The rationale of this method of task resolution is that, even if the tasks are abstract in our simulations, they semantically represent concrete ones: they involve a certain (controllable) amount of time; they are 'cumulative', in the sense that the total success depends on the success of its components; they can be achieved at different degrees (in our experiments the number of hits is used as a threshold of minimum performance and after being successful the agent can skip the other essays). Moreover, and most importantly for our theoretical model, the contribute of willingness (persistence) is clearly separated from ability; for each task 'attempting' is a prerequisite of 'doing' and an agent can fail in either. The contribution (positive or negative) of the environment is limited to the second phase, representing a favorable or unfavorable location for executing the task. The duration of the task is used to introduce another crucial factor that is *monitoring*: a delegator has to be able to control the activity of the delegee and possibly retire the delegation if it is performing badly; this aspect will be introduced in one of the experiments.

In the simulations we used *3 hits as a default*; however, all the effects are stable and do not depend on the number of hits. We have performed experiments with a different number of hits, obtaining similar results: choosing higher values leads to less tasks performed on average, but the effects remain the same.

This kind of task performing highlights both the role of ability and of willingness (persistence).

11.14.5 FCMs for Trust

For the sake of simplicity we have assumed that each cognitive agent has access to all true data of the other agents and of the environment; these data involve their task specific *ability*, their *willingness* and their current *environment*. All these data are useful for many strategies: for example, the *best ability trustor* always delegates to the agent with the higher ability value for that task.

The *cognitive trustor*, following our socio-cognitive model of trust, builds an elaborated mind model of all the agents; this is done with Fuzzy Cognitive Maps, as described in

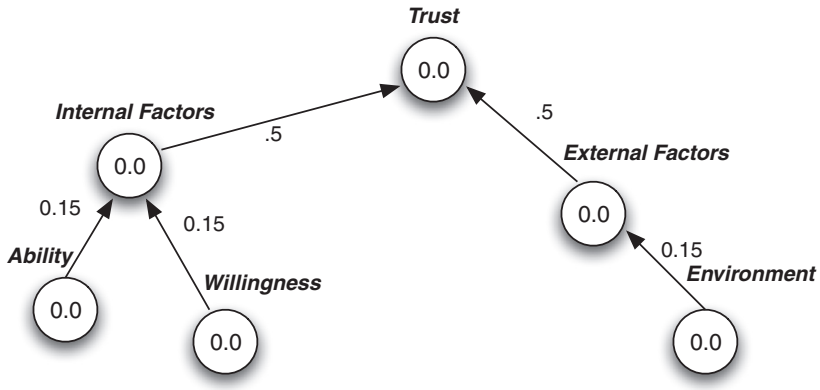

Figure 11.9 The FCM used by the *cognitive trustor*

(Falcone *et al.*, 2003).[7] The values of three nodes (ability and willingness as internal factors and environment as external factor) were set according to agent knowledge. The values of the edges reflect the impact of these factors and are always the same in the simulations. It has to be noticed that we never tried to optimize those factors: the results are always significant with different values. An additional point: while in the experiments the environment modifies the ability, in the 'mental representation' of FCMs this is not the case: this is not a piece of information that an agent is meant to know; what it knows is that there is an external (positive or negative) influence and it aggregates it by fulfilling the cognitive model. Figure 11.9 shows an (un-initialized) FCM.

11.14.6 Experiments Description

The first aim of our experiments is to compare the *cognitive trustor* and the *statistical trustor* in different situations: their delegation strategy represents two models of trust: derived from direct experience versus experience built upon a number of cognitive features. The random strategy was added as a baseline for the difficulty of the setting. The *best ability* and *best willingness* strategies are added in order to verify, in different settings, which are the single most influential factors; as it emerges from the experiments that their importance may vary, depending on some parameters.

In all our experiments we used exactly six agents (even if their delegation strategies may vary); it is important to always use the same number of agents, otherwise the different sets of experiments would not be comparable. In each experiment the task set by all agents is always the same; their ability set and willingness, as well as the environment influence, are randomly initialized.

The experiments are performed in a variable environment that influences (positively or negatively) the performance of some agents in some tasks, as previously explained.

[7] With respect to the general model, for the sake of simplicity we assume that unharmfulness and danger nodes are always 0, since these concepts have no semantic in our simulations.

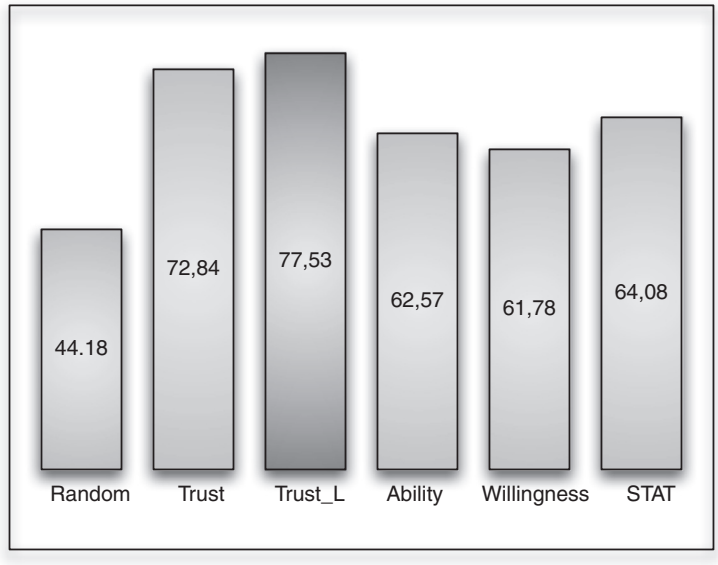

Figure 11.10 Experiment 1: comparison between many delegation strategies, 3 hits; (measuring the success rates). (Reproduced with kind permission of Springer Science+Business Media © 2005)

In order to allow the *statistical trustor* to learn from the experience, all the simulation sets were divided in two phases (two halves). The first phase is meant for training only: the statistical trustor delegates several times the tasks to all the agents and collects data from successful or unsuccessful performance. It uses these data in order to choose to whom to delegate in the second phase (in fact, it continues to learn even in the second phase). The delegation mechanism is always the same: it chooses the agent who has the best ratio between performed and delegated tasks; this number is updated after each result following a delegation. In order to measure the performance of this strategy, we analyzed only experimental data from the second phases.

The first experiment (***EXP1***) compares the *random trustor* (*RANDOM*), the best *ability trustor* (*ABILITY*), the *best willingness trustor* (*WILLINGNESS*), the *statistical trustor* (*STAT*), and two other cognitive strategies that differ only because of how much they weight the environmental factor: *no impact* (*TRUST*) does not consider the environment, while *low impact* (*TRUST_L*) gives it a low impact (comparable to the other factors). In Figure 11.10 we show 250 simulations for 100 tasks.

We can see that the cognitive strategies always beat the statistical one.[8] Moreover, it is important to notice that recognizing and modeling the *external components* of trust (the environment) leads to a very high performance: the cognitive trustor who does not consider the environment (*TRUST*) beats the statistical one (*STAT*), but performs worse than the cognitive trustor who gives a role to the environment (*TRUST_L*).

[8] The results are similar e.g. with five hits (250 simulations, 100 tasks): *RANDOM*: 26,24; *TRUST*: 57,08; *TRUST_L*: 61,43; *ABILITY*: 40,58; *WILLINGNESS*: 48,0; *STAT*: 49,86.

We have performed three more experiments in order to verify another interesting condition about learning (250 simulations, 100 tasks). Sometimes it is not possible to learn data in the same environment where they should be applied. For this reason, we have tested the *statistical trustor* by letting them learn without environment and applying data in a normal environment (***EXP2*** – positive and negative influences as usual), in an always positive environment (***EXP3*** – only positive influences), and in an always negative environment (***EXP4*** – always negative influences). As easily foreseeable, the mean performance increases in an always positive environment and decreases in an always negative environment; while this is true for all strategies, the statistical strategy has more troubles in difficult environments. Figure 11.11

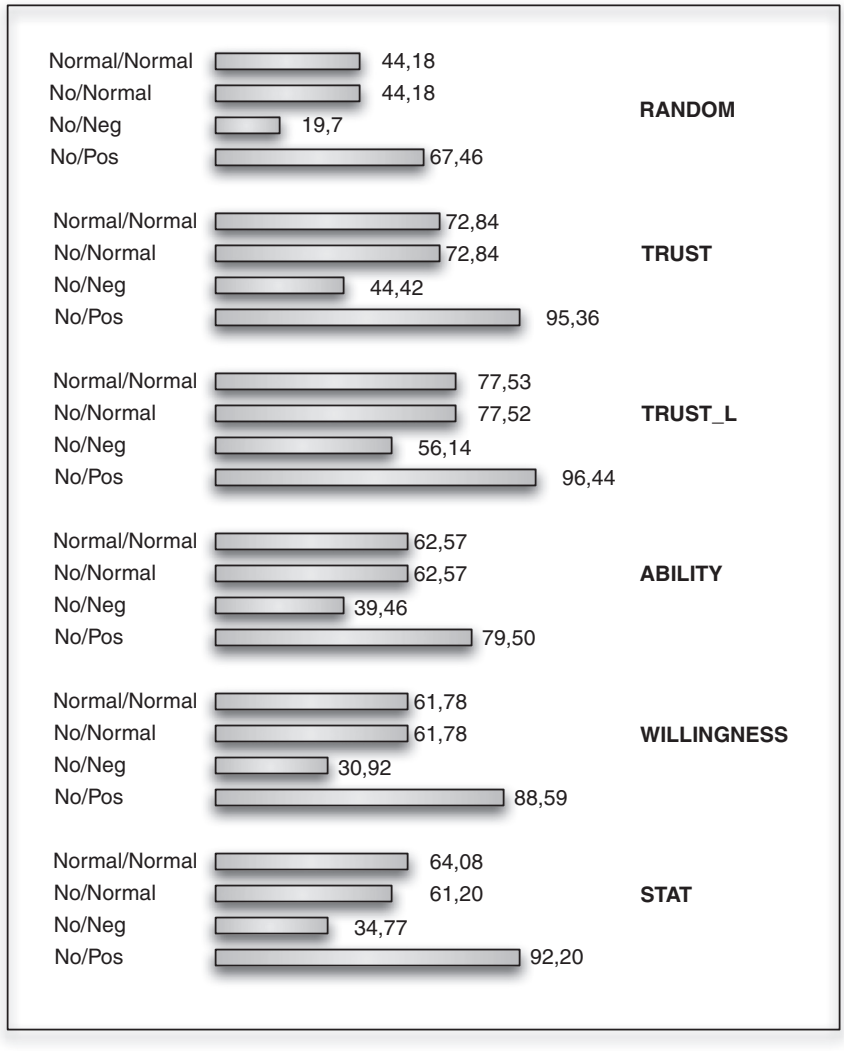

Figure 11.11 Experiments 1, 2, 3 and 4 compared (measuring the success rates). (Reproduced with kind permission of Springer Science+Business Media © 2005)

shows four cases:

1) learning in normal environment, task in normal environment (normal/normal);
2) learning without environment, task in normal environment (no/normal);
3) learning without environment, task in negative environment (no/neg);
4) learning without environment, task in positive environment (no/pos).

11.14.7 Using Partial Knowledge: the Strength of a Cognitive Analysis

The results achieved in the above experiments are quite interesting, but rather predictable. More interesting and with high degree of difficulty of prediction is the experiment in which we try to individuate the level of approximation in the knowledge of a cognitive trustor about both the properties of other agents and of the environment. In other words, we would like give an answer to the questions: *when is it better to perform as a cognitive trustor with respect to a statistical trustor? What level of approximation in the a priori knowledge is necessary in order that this kind of trustor will have the best performance?*

To answer this interesting question we have made some other experiments (as ***EXP5***) about errors in evaluation. As already stated, all the values we assume about cognitive features are true values: each agent knows all the real features of the others.

This is an ideal situation that is rarely implemented in the real world/system. In particular, in a multi-agent system environment there can be an evaluation process that is prone to errors. In order to evaluate how much error the cognitive trustor can deal with without suffering from big performance losses, we have compared many cognitive trustors introducing some different levels of 'noise' in their data.

Figure 11.12 shows the data for the *random trustor (RANDOM); the best willingness trustor (WILLINGNESS); the best ability trustor (ABILITY)*; the normal *cognitive trustor (NOERR)*, as well as some other *cognitive trustors (ERR_40, ERR_20*, ...) with 40%, 30%, 20%, 10%, 5%, 2.5% error; the *statistical trustor* (STAT). While all the experiments used a set of six agents, we present aggregated data of different experiments. We have ordered the strategies depending on their performance; it is easy to see that even the worst cognitive *trustor* (40% error) beats the statistical trustor. Under this threshold we have worse performances.

Real Time Experiments

We have performed some real time experiments, too. ***EXP6*** (see Figure 11.13) involves three cognitive strategies in a normal environment (250 simulations, 500 tasks).

The differences between the cognitive trustor without environment and the two with environment are statistically meaningful; the difference between the two cognitive trustors with environment are not. The results are very close to those that use turns; the differences depend on the limited amount of time we set for performing all the tasks: by augmenting this parameter more quickly, strategies become more performing.

Another experiment (***EXP7***) aims at testing the performance of the first trustworthy trustor (***FIRST***). Here there are two parameters for performance: *Credits* and *Time*. *Time* represents

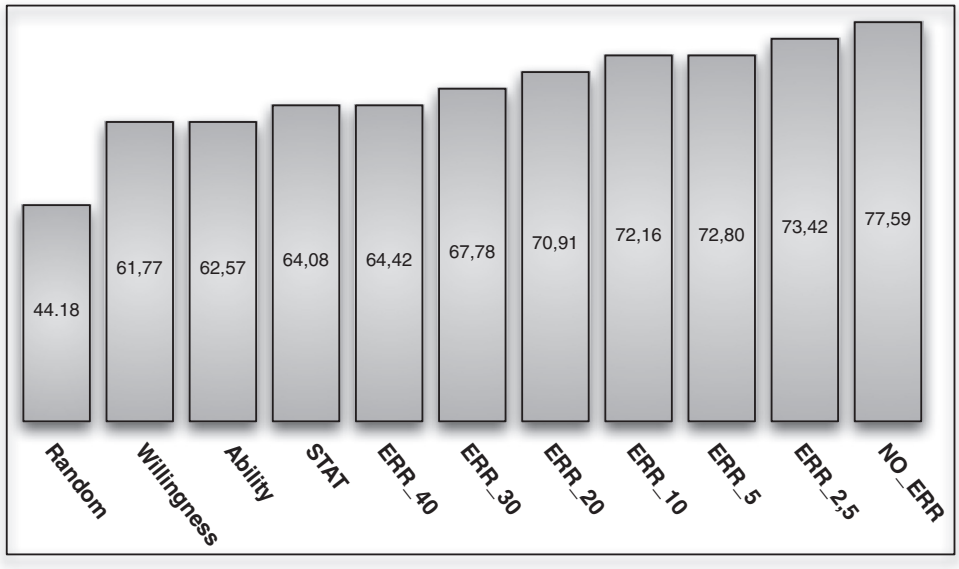

Figure 11.12 Experiment 5: introducing noise (measuring the success rates). (Reproduced with kind permission of Springer Science+Business Media © 2005)

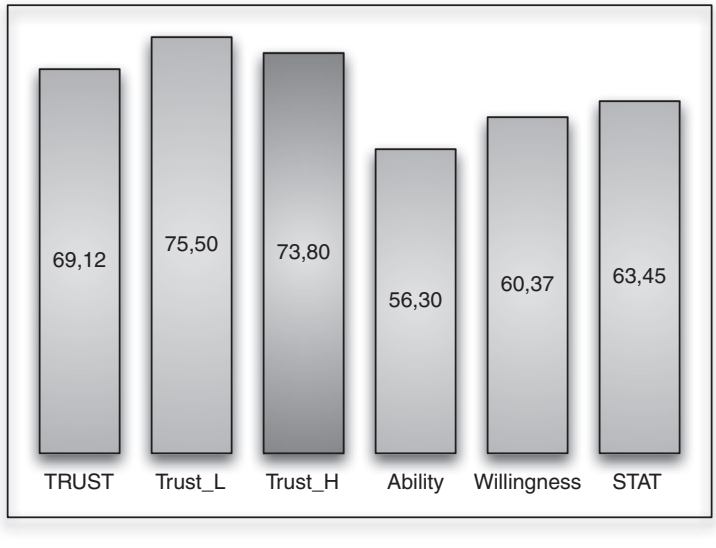

Figure 11.13 Experiment 6: real time (measuring the success rates). (Reproduced with kind permission of Springer Science+Business Media © 2005)

how much time is spent in analyzing offers and delegating, i.e. how much offers an agent collects before choosing.[9]

While in the preceding experiments the agents collected all the offers before deciding, here an agent can delegate when it wants, saving time. This situation is closer to a real MAS situation, where agents act in real time and sometimes do not even know how many agents will offer help. How much time is spent in delegation depends on the strategy and on simulation constraints. The random trustor can always choose the first offer it has, so it results in being the quickest in all cases. If there is a fixed number of agents and the guarantee that all of them will offer, best ability, best willingness, the cognitive trustors and the statistical trustor can build and use an ordered list of the agents: so they have to wait until the offer from the pre-selected agent arrives. In the more interesting MAS scenario, without a fixed number of offering agents, each incoming offer has to be analyzed and compared with the others. In order to avoid waiting ad infinitum, a maximum number of offers (or a maximum time) has to be set.

However, in this scenario there can be other interesting strategies, such as the first trustful trustor, who does not wait until all six offers are collected, but delegates when the first 'good offer' (over a certain threshold) is met; this can lead to more or less time saved, depending on the threshold. Here we present the results of ***EXP7*** (250 simulations, 100 tasks); in this case all agents wait for exactly six offers (and compare them) before delegating, except for the random trustor (who always delegates to the first one) and the first trustful trustor who delegates to the first one that is over a fixed threshold. Figure 11.14 and Figure 11.15 show the results for credits (as usual) and time spent (analyzed offers).

The first trustworthy trustor still performs better than the statistical trustor, saving a lot of time. Depending on the situation, there can be many ways of aggregating data about credits and time. For example, in a limited time situation agents will privilege quickness over accurateness; on the contrary, in a situation with many agents and no particular constraints over time it would be better to take a larger amount of time before delegating.

Experiments with Costs

In our simulations we assume that the costs (for delegation, for performing tasks, etc.) are always the same; in the future it would be interesting to introduce explicit 'costs' for operations, in order to better model real world situations (e.g. higher costs for more skilled agents).

The costs are introduced as follows. When an agent sends their 'proposal' they quote a price that averages to their ability and willingness (from 0 to 90 Credits).[10] Half the cost is paid on delegation, the second half is paid only if the task is successfully performed. If a task is re-delegated, the delegator has to pay again to delegat. For each successfully performed task the delegator gains 100 credits. There may in the future even be costs for receiving the reports, but at the moment this does not happen.

In order to model an agent who delegates taking into account the gain (and not the number of achieved tasks) we have used a very simple utility function, that simply multiplies trust and (potential) gains minus costs. It has to be noticed that on average a better agent has a

[9] There are other possible parameters, such as time spent in reasoning or in performing a task. However, we have chosen only the parameter which is more related to the Delegation Strategy; the other ones are assumed to have fixed values.

[10] We generate randomized values over a bell curve which averages it (ability + willingness)/2.

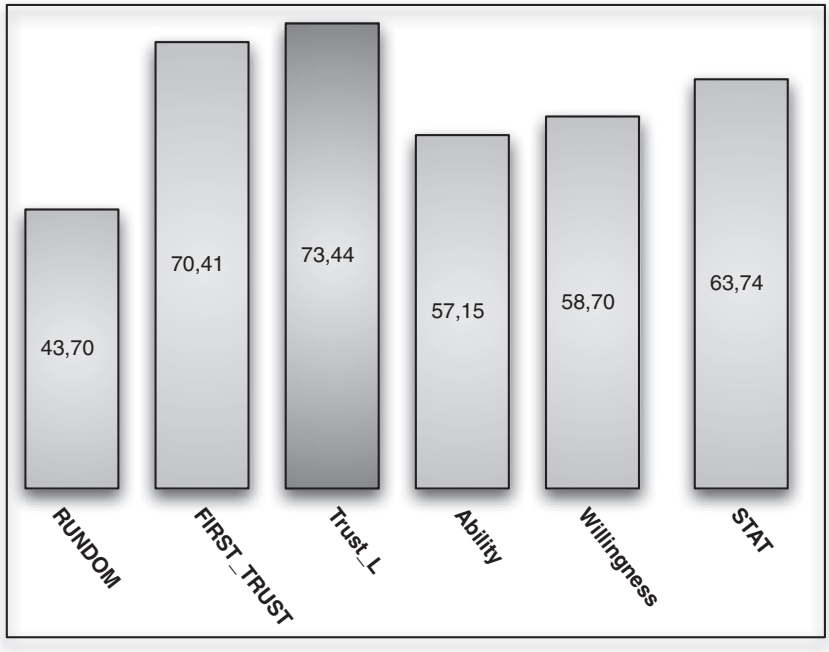

Figure 11.14 Experiment 7: real time, introducing the first trustworthy strategy (measuring Credits). (Reproduced with kind permission of Springer Science+Business Media © 2005)

higher cost; so the agent who maximizes trust is penalized with respect to gain. The agent who maximizes following the utility function chooses agents with less ability and willingness; for this reason it performs fewer tasks than the other agents.[11]

EXP8 (see Figure 11.16) was performed for 200 tasks and 250 simulations. The results refer (1) to the tasks performed;[12] (2) to the gains; (3) to time spent. The policies are: *random*; *cost_trust* (that uses the utility function); *trust_ambient* (that uses trust as usual); *low_cost* (that always delegates to the less expensive); *first_trust* (that delegates to the first over a certain trust threshold), *stat* (that performs statistical learning).

The most important result is about gains; the agent who explicitly maximizes them has a great advantage; the trust strategies perform better than the statistical one, even if they all do not use information about gains but about tasks achieved. It is even interesting to notice that the worst strategy results delegating always to the least expensive: in this case costs are related to performance (even if in an indirect way) and this means that cheap agents are normally bad.

Delegation and Monitoring

In order to model a more complex contract net scenario, we have introduced the possibility of the delegator monitoring the performance of the delegated agents into the intermediate steps

[11] We have chosen a little difference between costs and gains in order to keep costs significant: setting lower costs (e.g. 1–10) makes them irrelevant; and the trust agent performs better.

[12] Multiplied * 10 in order to show them more clearly.

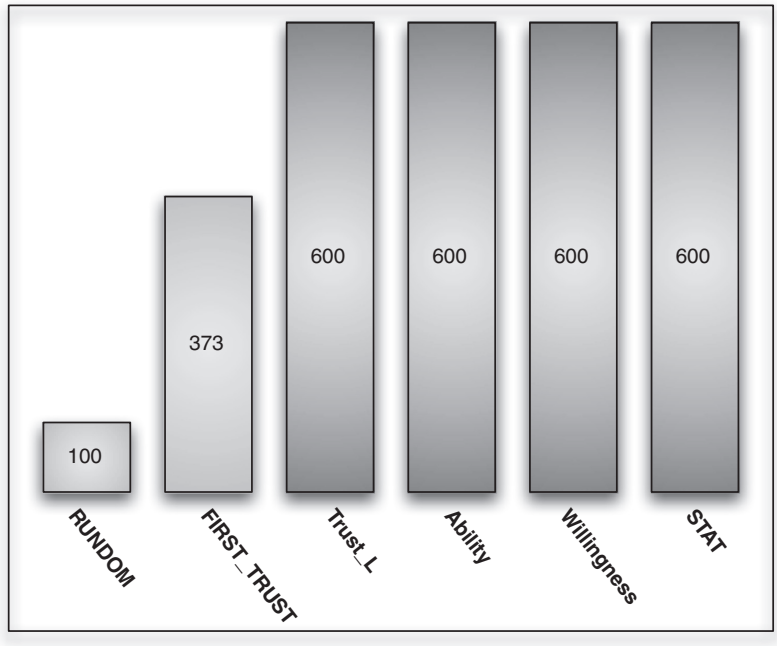

Figure 11.15 Experiment 7: real time, introducing the first trustworthy strategy (measuring time spent). (Reproduced with kind permission of Springer Science+Business Media © 2005)

of the task; he can decide for example to stop the delegation and to change delegee. In this case, the agent has a new attribute: *controllability*. A more complex task resolution scheme and cost model (e.g. the costs of stopping or changing the delegation) is needed, too.

In order to exploit the controllability, we changed the task resolution scheme. In order to be successful, an agent has, as usual, to *score a certain number of hits* (e.g. three). In order to score a hit, instead of performing a number of attempts equal to their willingness, the agent has *exactly 10 attempts*: for each attempt, he first tests his willingness (i.e. if it actually tries to perform it); if this test is successful, it tests his ability (i.e. if the delegee is able to perform the task). Each test is simply 'rolling a random real number in (0,1): if the roll is less than the tested attribute (Willingness or Ability), the test is successful. The environment can interfere with the agent's activity setting the Difficulty of each Ability test. At the end of the ten attempts, if the number of scored hits (e.g. three) is sufficient, the task is performed; otherwise it is not.

In addition to the usual tests of willingness and ability, for each one of the ten attempts each Agent checks if the delegee sends a report (a *controllability* check); the report contains the number of the current attempt, information about the activity in this attempt (tried/not tried; performed/not performed) as well as the number of successes achieved. At the end of the ten attempts, the agent sends a final message that says whether the task was successful or not.

Monitoring consists of receiving and analyzing the reports. Basing on this information, an agent can decide either to confirm or to retire the delegation: this can only be done before the final message arrives. If they retire the delegation, they can repost it (only one more time). Obviously, agents having a higher controllability send more reports and give more opportunities to be controlled and stopped (and in this case the tasks can be reposted).

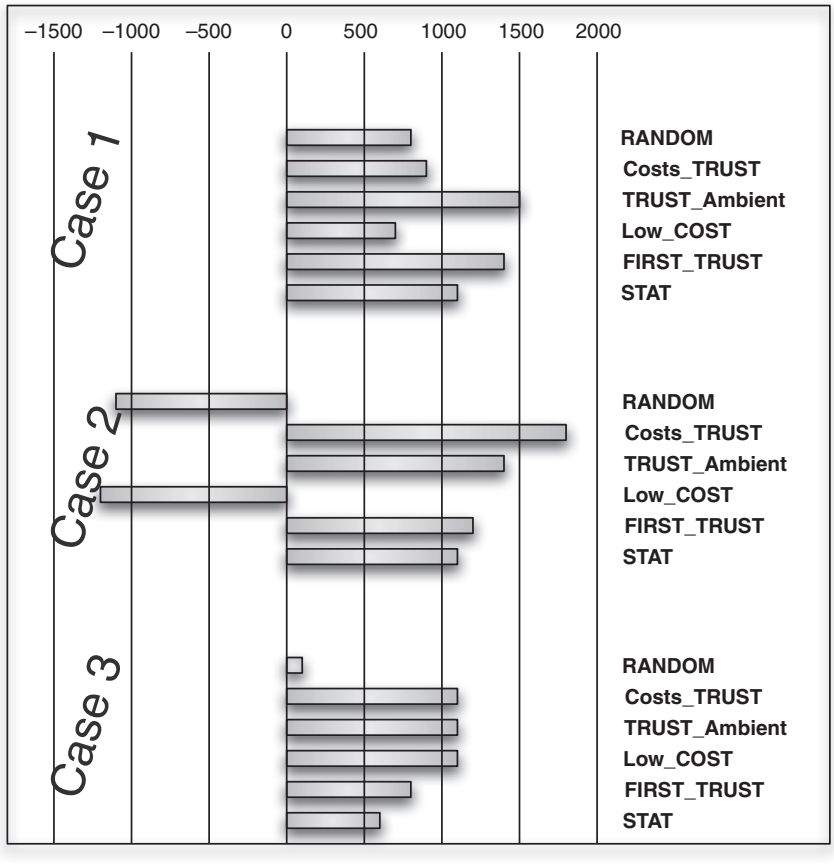

Figure 11.16 Experiment 8: comparing different delegation strategies in different cases: number of tasks performed (success rates, case 1), number of credits (gains, case 2), and number of analyzed offers (time spent, case 3). (Reproduced with kind permission of Springer Science+Business Media © 2005)

In ***EXP9*** (see Figure 11.17) we use controllability and we offer the possibility of re-posting a task. An agent who uses controllability as an additional parameter for trust (i.e. giving a non null weight to the corresponding edge in its FCM) is compared with agents who do not use it; it is (more or less) biased towards choosing agents who give more reports and so they have more possibilities of re-posting unsuccessful tasks (tasks can be reposted only once). All agents (except ***TRUST_BASE***) use the following heuristic in order to decide whether to retire delegation:

$$if\ (10 - current\_essay/current\_essay < 1)\ and\ (success\_number < 3)$$
$$and\ (I\_can\_repost)\ then\ retire\_delegation$$

The experiment was performed for 200 tasks and 250 simulations. ***RANDOM*** is the baseline; ***TRUST*** is the agent who uses ability, willingness and environment but without controllability; ***TRUST_BASE*** is the same agent but the one who never reposts their tasks; ***CONTROL_1***

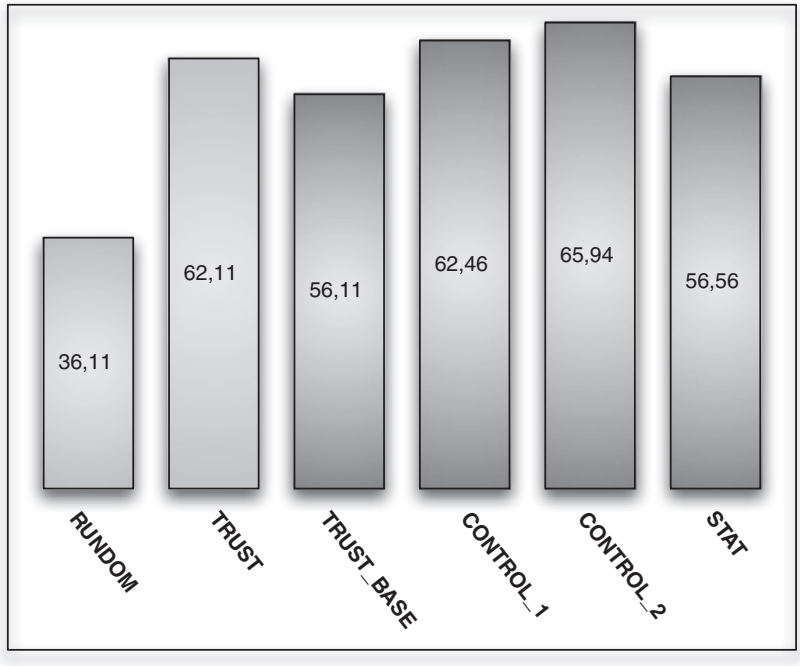

Figure 11.17 Experiment 9: monitoring (measuring the success rates). (Reproduced with kind permission of Springer Science+Business Media © 2005)

introduces a little weight on controllability; ***CONTROL_2*** introduces a significant weight on controllability; ***STAT*** is statistical.

Note that since the task resolution scheme is changed, these results are (on average) lower than in the other experiments; however, it is possible as usual to compare the different strategies. Considering controllability gives a significant advantage over the other strategies.

11.14.8 Results Discussion

In our experiments we have tried to compare different trust strategies to delegate tasks in a contract net. The setting abstracts a simplified real-world interaction, where different agents have different capabilities (mostly represented by ability) and use more or less similar resources (mostly represented by willingness) in order to realize a certain task. The role of the environment is also very relevant because external conditions can make the tasks more or less easy to perform. On the basis of their trust, the delegating agents (with different strategies) decide to whom to assign their tasks.

We analyzed two concepts of trust:

- the first (referred to as the *statistical trustor*), that it is possible to model the trustworthiness of other agents only on the basis of the direct (positive or negative) experience with them; on the fact that there is only one dimension to consider: the number of successes or failures the agents preformed in the previous experiences.

- the second (referred to as the *cognitive trustor*), based on a more complex set of factors that have to be considered before trusting an agent; in particular, both a set of trustor features and environmental features.

In all our experiments the cognitive trustors perform better than the statistical ones, both from the point of view of global successes (number of *Credits*) and stability of behavior (less standard deviation in the simulations data). The cognitive strategy models the agents and environment characteristics more effectively and it allows the resources to be allocated in a highly accurate way. Introducing a changeable environment does not decrease performance, providing that it is considered as a parameter; but even if it is not considered, the results are largely better than with a statistical trustor.

The fact that an algorithm that knows the real processes implemented by the agents when achieving tasks uses a simulation mechanism of these processes for selecting the best performances is quite predictable. For this reason we have made new experiments introducing a significant amount of noise in the cognitive agent knowledge. The results show that the performance of the cognitive agent remains better than the statistical one up to an error of 40%. So, the cognitive trustor is very accurate and stable under many experimental conditions. On the contrary, even with a large amount of data from learning (the training phase), the statistical strategy is not performing well. Moreover, if the learning is done in a different environment, or if the environment is particularly negative, the results are even worse.

With a low number of hits (e.g. three) the task is designed to privilege ability over willingness; however, augmenting the number of hits, the relative relevance changes. A strong environmental influence shifts the equilibrium, too: it modifies the ability scores which become more variable and less reliable. Modifying the relative weight of those parameters (depending on the situation) into the FCM of the cognitive trustor can lead to an even better performance.

In the real time experiments, when time is implicitly introduced as an additional performance measure, a variant of the *cognitive trustor*, the *first trustful trustor*, becomes interesting: it maintains high task performance (measured by credits) with a limited amount of time lost.

Introducing costs into the experiment leads the agents to maximize another parameter, gain, with respect to tasks achieved. It is not always the case that more tasks mean more gain, because many agents who perform well are very costly; in fact the best strategy optimizes gains but not the number of achieved tasks.

Introducing a monitoring strategy, with the possibility of retiring the delegation and reposting the task, introduces an extra possibility for the agents, but also another difficulty, because each re-post is costly. Considering explicitly the controllability as a parameter for trusting an agent gives a significant advantage, because more controllable agents – especially in real time – enable a more precise distribution of the tasks and even a recovery from wrong choices.

Depending on the situation (e.g. with or without environment; with or without the possibility of retiring the delegation) and from the goals (e.g. maximize tasks, time or gains) the possible delegation strategies are many. In all cases, trust involves an explicit and elaborated evaluation of the current scenario and of the involved components – and our results demonstrate that this gives a significant advantage in terms of performance with respect to a mono-dimensional statistical strategy.

11.14.9 Comparison with Other Existing Models and Conclusions

Many existing trust models are focused on reputation, including how trust propagates into recommendation networks (Jonker and Treur, 1999) (Barber and Kim, 2000) (Jøsang and Ismail, 2002). On the contrary, our model evaluates trust in terms of beliefs about the trustee's features (ability, willingness, etc.); reputation is only one kind of source for building those beliefs (other kinds of source are direct experience and reasoning). In the present experimental setting there is not any reputational mechanism (that we could also simulate in the cognitive modeling), so a comparison with these models is not appropriate.

There are some other approaches where trust is analyzed in terms of different parts; they offer a more concrete possibility for comparison. For example, in (Marsh, 1994) trust is split into: *Basic Trust*, *General Trust* in agents, *Situational Trust* in agents. Basic trust is the general attitude of an agent to trust other agents; it could be related to our model if it is considered as a general attitude to delegate tasks to other agents in the trust relationships; in the experiments already illustrated, we did not consider the possibility of introducing agents with the inclination to delegate to others or to do the task themselves. In any case, the setting can certainly include these possibilities. General trust is more related to a generic attitude towards a certain other agent; the more obvious candidate in our setting is willingness, even if the two concepts overlap only partially. Situational trust is related to some specific circumstances (including costs and utilities, that are not investigated here); there is a partial overlap with the concept of ability, that represents how well an agent behaves with respect to a certain task. So, the model presented in (Marsh, 1994) is, to a certain extent, comparable with our one; however, it lacks any role for the environment (more in general for the external conditions) and it introduces into trust the dimensions of costs and utility that in (Castelfranchi and Falcone, 1998), (Falcone and Castelfranchi, 2001) are a successive step of the delegation process that is presented in a simplified way.

Our experiments show that an accurate socio-cognitive model of trust allows agents in a contract net to delegate their tasks in a successful way. In any case, for better testing the model it is be necessary to realize a set of new experiments in which we even allow the *cognitive trustors* to learn from experience. While the learning of the *statistical trustor* is undifferentiated, the cognitive trustor is able to learn in different ways from different sources. In the model for each trust feature there are (at least) four different sources: direct experience, categorization, reasoning, reputation; each of them contributes in a different way. More, higher level strategies can be acquired: for example, depending on the environment and the task difficulty (number of hits) an optimal weight configuration for the FCMs can be learned.

Other directions of work could be to experiment with agents starting with some a priori knowledge about other agent's stats with a percentage of error, and in which they can refine this percent by analyzing how well they perform in the delegated tasks. This is a kind of statistical, not specific learning. However, in order to learn in a more systematic way, an agent has to discriminate each single stat. In order to do this, they could analyze the incoming reports (e.g. how many times an agent tries a task for willingness; how many times they perform it for ability; how many reports they send for controllability). The controllability stat introduces an upper limit even to how many learning elements an agent can receive, so it becomes even more critical.

Finally, we would like to suggest that a reputation and recommendation mechanism is included, in order to add another trust dimension to the simulations. In this way we could

introduce new delegation strategies (based on the existing systems in literature), and study how reputation interacts with the other cognitive features in the cognitive trustor.

References

Barber, S., and Kim, J. (2000) Belief revision process based on trust: agents evaluating reputation of information sources, *Autonomous Agents 2000 Workshop on 'Deception, Fraud and Trust in Agent Societies'*, Barcelona, Spain, June 4, pp. 15–26.

Castelfranchi, C. Reasons: belief support and goal dynamics. *Mathware & Soft Computing*, 3: 233–247, 1996.

Castelfranchi, C. and Falcone, R. Towards a theory of delegation for agent-based systems, *Robotics and Autonomous Systems*, Special issue on Multi-Agent Rationality, Elsevier Editor, 24 (3-4): 141–157, 1998.

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

Dragoni, A. F. (1992) A model for Belief Revision in a Multi-Agent Environment. In *Decentralized AI - 3*, Y. Demazeau, E. Werner (eds.), pp. 215–231. Amsterdam: Elsevier.

Dubois, D. and Prade, H. (1980) *Fuzzy Sets and Systems: Theory and Applications*, Academic Press, Orlando, FL.

Falcone, R. and Castelfranchi, C. (2001) Social trust: a cognitive approach, in C. Castelfranchi and Y. Tan (eds.), *Trust and Deception in Virtual Societies*, Kluwer Academic Publishers, pp. 55–90.

Falcone, R., Pezzulo, G., Castelfranchi, C. A fuzzy approach to a belief-based trust computation. *Lecture Notes on Artificial Intelligence*, 2631, 73–86, 2003.

Falcone, R., Pezzulo, G., Castelfranchi, C. (2005) A fuzzy approach to a belief-based trust computation, in Falcone, R., Barber, S., Sabater-Mir, J., Singh, M., (eds.), *Trusting Agents for Trusting Electronic Societies*. Lecture Notes on Artificial Intelligence, n°3577, Sprinter (pp. 43–58).

Jonker, C., and Treur, J. (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.

Jøsang, A., and Ismail, R. (2002) *The Beta Reputation System*. In the proceedings of the 15th Bled Conference on Electronic Commerce, Bled, Slovenia, 17–19 June 2002.

Kosko, B. (1986) Fuzzy cognitive maps. *International Journal Man-Machine Studies*, 24: 65–75.

Marsh, S.P. (1994) Formalising trust as a computational concept. PhD thesis, University of Stirling. Available at: http://www.nr.no/abie/papers/TR133.pdf.

Pezzulo, G. and Calvi, G. (2004) AKIRA: a Framework for MABS. *Proc. of MAMABS 2004*.

Schillo, M., Funk, P., and Rovatsos, M. (1999) Who can you trust: Dealing with deception, Autonomous Agents '99 Workshop on 'Deception, Fraud and Trust in Agent Societies', Seattle, USA, May 1. 1: 81–94.

Smith, R.G. (1980) The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*.